Assessing Lexical Accessibility: A Critical Review of Three Extant Tests and a New Approach.

David Coulson

Key Words: Word Recognition; Lexical Competence; Test Validity; Consistency; Rasch Analysis

Abstract

Vocabulary knowledge is the basis of language ability. For learners, increasing the number of words and phrases they know is the most important thing they can do to improve their ability. However, lexical competence involves much more than simply memorizing long lists. All words have various aspects of knowledge. Some examples are syntactic and semantic behavior, derivations, active and passive knowledge, collocations and the network of associations and knowledge of polysemy. (e.g. Richards, 1976; Melka, 1997) However, for a typical inter-mediate learner who knows several thousand head words, assessing the overall state of her vocabulary knowledge would be impossible if all these aspects were taken into account. Understandably, teachers often rely on one-time vocabulary tests of L2-L1 translation knowledge. But since such tests constitute only a tiny fraction of known words and target only one facet of knowledge, they reveal little about the overall state of lexical development. Moreover, the development of lexical competence is not linear or monotonic. It progresses in spurts and regresses with disuse, as any committed language learner has experienced. Clearly, sensitive, practical testing tools are needed to assess the various facets of L2 lexical ability. This paper will deal with the testing of L2 word recognition ability.

Introduction

There has not been much research on the question of how teachers can assess their learners' competence beyond the limitations of using simple translation tests. Practical tests for size (e.g. Nation, 1983; Meara & Buxton, 1990) and associational knowledge (e.g. Wesche & Paribakht, 1996) and V-Links (Meara and Wolter, 2004) have been proposed and actively researched in recent years. In recent years, the importance of orthographical knowledge of words as one aspect of lexical competence has greatly increased. Research has shown that successful reading is based on bottom-up sampling of each word in text, and not on a top-down mode of processing where the eyes of the reader jump over words which can be guessed. (Rayner & Balota, 1989) During each saccade in reading (the duration that the eye focuses on each part of the text), the actual time taken up by lexical recognition is about 50 milliseconds in native speakers. This time is longer in non-native speakers but becomes increasingly rapid with practice and language processing in general. The practical problem concerning assessment that this presents is that latencies in the order of 50 milliseconds (or the even smaller scale changes that characterize L2

英文学科

word recognition development) are impossible to measure without specialist equipment.

It is this aspect of the development of L2 word-recognition latencies, and their measurement, that I address in this paper. In Part 1 of the paper, I will critically review three papers (Shiotsu, 2002; Jacobsen, 1995; Laufer and Nation 2001) which deal with lexical access in three distinct ways. In Part 2, I will describe my research on a test of word recognition speed, 'Q_Lex'. It aims to provide an overall measure of the change of reaction times to basic word stimuli. Q_Lex purposely measures reaction times to only high frequency vocabulary. With general improvement in L2 ability, the reaction times to basic words should become increasingly automatic, as learners progress through intermediate standard and on. Q_Lex tracks this development by an increase in the number of words found within the norms of native speaker performance. It is envisaged that Q_Lex will eventually be used in conjunction with other global measures such as size and associational knowledge. However, for now, I will present basic results which demonstrate that recognition speed develops in parallel with general L2 development. I will also show results that establish reliability by examining scores over two tests.

Part 1
A Critical Review of Important Research papers in this field.
Critical Review 1: Shiotsu, 2002.

In this paper, Shiotsu replicates, in part, an experiment on the relationship between word-decoding skill and reading ability by Haynes (1989). The original investigation was conducted on Taiwanese and American students. Shiotsu investigates the performance of Japanese students only. To test decoding ability, Haynes gave subjects a long list of lexical decision tasks printed on paper. These consisted of pairs of stimuli 4 letters long. Subjects had to judge if each pair was identical or different. Four different types of stimuli were created. The first kind used real words, the second pseudo-words (e.g. 'gane'), and the third illegal letter strings (e.g. 'gvae'). Here orthographic decoding speed was at issue. Shiotsu refers to these tests as tapping surface recognition ability. The fourth kind of stimuli pair involved judging if a pair of words has the same or different meaning, which is posited to require deeper semantic access. Haynes reported that Taiwanese EFL students were significantly slower on pseudo-words than real words, and even slower on illegal strings. In contrast, native speakers reacted in similar times to real words and pseudo-words, and were significantly slower only on reaction times to illegal letter strings. Further, it is reported that these learners' passage reading speed was significantly related to the word matching tasks. Based on these findings, Shiotsu reports Haynes' claim that native speakers are able to decode and analyze words more fluently than EFL learners.

Shiotsu adapted these tests to a computer format to investigate the 'visual processing efficiency' of Japanese university students. The reason for adapting the test to a computer-based measurement was, principally, to collect the latency information for each pair of stimuli. Shiotsu explains his reasoning by saying that a paper-based test is only capable of recording the performance of subjects on the test as a whole. Inexplicably, however, the author fails to make any use of the latency measure of stimuli in the test. Finally, Shiotsu particularly wanted to know if good readers are faster at decoding alphabetic stimuli than are poor readers. To this end, he gave his subjects a reading test and compared the results of this against their performance on the four kinds of lexical-decision tasks. He reports a strong correlation between

the strongest readers and higher scores on the tasks, especially for the synonym/antonym task which requires semantic access.

Summary

Shiotsu starts with the hypothesis that many Japanese EFL readers may have 'underdeveloped decoding fluency' in reading English. With regard to this, he considers the issue of whether their passage reading speed is related to speed of visually processing individual words an empirical one. His study comprises three aims:

1) An investigation of the reliability of the word recognition measurement tool. These are the four kinds of lexical decision tasks outlined above.

2) An investigation about whether, and how, Japanese EFL readers are affected when the target stimuli are pseudo-words, or illegal letter strings.

3) An investigation whether good text comprehenders are faster at recognizing the words in the tasks.

In her 1989 study, Haynes made use of 48 pairs of stimuli for each of the four kinds of tasks, written out on paper. (words, pseudo-words, illegal letter strings and synonym and antonym pairs, shortened to 'S/A') The task was to judge if they were the same or different by circling 'S' or 'D'. An example would be the following:

care	card	S	D	(*word*)
bule	bele	S	D	(*pseudo-word*)
botp	botp	S	D	(*illegal letter string*)
fail	pass	S	D	(*synonym/antonym pair*)

The items were high frequency words with a mean occurrence of 670 per million. The four kinds of tests were programmed to appear, one at a time, on a personal computer. Students clicked on a key to indicate if they know the word, and this was timed in milliseconds. In total, the test required between 10 and 20 minutes to complete.

Shiotsu developed the format of judging students on their mean times, adapting it to a computer-based test in which each stimulus would appear one at a time. 12 items were derived from Haynes' study for the first three kinds of test (all 4 letters). Shiotsu claims that only 12 of each kind were necessary to obtain a reliable estimate of student performance, since individual items' mean reaction times could be measured accurately by the programme. As for the synonym/antonym pairs, 36 were chosen from Haynes' 48 (A mean 4.89 letters varying from 3 to 9 letters in length.) Testees were to push 'S' or 'D' on a keyboard. Since latency information was obtained for each pair, Shiotsu decided to reduce the number of items from 48 to 12 for each category.

Japanese university students were divided into 2 groups, based on above or below average performance on a reading test. There were 4 short passages of around 170 words which required students to answer comprehension questions pertaining to the overall meaning of the text. 20 questions were carefully prepared so that they were not simply a search of verbatim information or could be answered by paraphrasing information.

Results for the 67 students who completed all tests are summarized in table 1.

Concerning the first question of reliability, S/A items were highest at .84, and pseudo-words

Table 1. Descriptive Statistics (N = 67)

	k	Min	Max	Mean	SD	Reliability
Reading	20	1	19	12.25	4.72	.85
Word	12	653	1369	922.08	155.18	.74
Pseu	12	701	1506	1004.75	179.92	.65
Irrg.	12	705	2023	1151.45	260.84	.75
S/A	36	950	2502	1564.52	362.17	.84

the lowest at .65. The second question asked if Japanese EFL readers are affected when the target stimuli are pseudo-words, or illegal letter strings. A repeated measures ANOVA revealed a significant effect of the stimulus type F (1.443, 95.23)=150.828, $p<0.001$.

By extension, Shiotsu claims that the second question of whether pseudo-words or illegal letter strings affect Japanese students is elucidated. Like the Taiwanese subjects in Haynes' study, here too the Japanese students were slower at responding to pseudo-words than real words and slower again at pseudo-words.

Concerning the third question of whether good readers are faster at recognizing words, Shiotsu divided his group into two, those above average and those below, and considered their results on the lexical decisions tasks. The results are shown in Table 2.

In particular the difference in reaction time to the synonym/antonym task is pronounced, with stronger readers showing markedly faster judgment time. The results for the other tasks were much less clear-cut. There clearest difference between the readers was on the word decision task although the gap was only 51 milliseconds between the two groups. On the irregular string decision task, there was a similar gap of 47 milliseconds, but this was in favour of the weaker readers.

Table 2: Latency by Group

	Above Average N=37	Below Average N=30
Real Words	900.19 (153.03)	949.08 (156.11)
Pseudo-words	1002.78 (203.01)	1007.18 (150.02)
Irregular Strings	1172.56 (258.12)	1125.42 (266.09)
S/A	1435.01 (297.09)	1724.25 (375.71)

SD in brackets

Critique

Shiotsu provides reasonable evidence that a lexical test of synonym/antonym recognition is

more closely linked with passage reading ability than easier same-or-different decision tasks. However, several aspects of Shiotsu's experimental procedure can be questioned. To put his findings in perspective, I will present the results of my own small replication of a similar set of lexical decision tasks which prove to be at odds with some of the findings reported above. Further, Shiotsu's findings are not consistent with separate research by Haynes and Carr, published a year after the work replicated here. All told, this leads me to suspect that Shiotsu's findings are compromised by his methodology.

The first point on which Shiotsu can be criticized is the title of his work. Rather than elucidating the doubtless complex issue of individual differences in L2 recognition speed in his study, as promised in the title, he simply reports group means. This is an odd situation since it was to elucidate individual differences, and latencies to each test item, that Shiotsu recreated Haynes' study using a computer in the first place. In fact, Shiotsu provides no compelling reason why this research needed to be done on computer. He could have come to the same conclusions by using paper-based tests, as Haynes did. In fact, he may have compromised his findings. One can imagine, for example, that since testees were allowed to relax between each item and proceed to the next when they were ready, there was a qualitative difference in testing conditions from Haynes' study, in which her subjects rushed down the lists of items on the page as quickly as possible. Although latencies for each type are reported, we do not know how well the subjects might have performed in comparison to native speakers, or against an imposed time limit. This would have easily provided some measure of individual variability.

Second, it is odd that Shiotsu would choose to reduce the number of items in his test from the original 48 of each type to only 12 (36 in the case of the S/A items.) The criteria for their selection are not described. He claims he did this because latency information for individuals could be measured accurately by computer (although he does not actually report this). This is a very small sample size to assess performance on the three kinds of stimuli. One can imagine this small sample size would not be reliable. Evidence for this comes from the results for irregular strings. Below average readers performed more strongly on this category (1125.42 msecs) than those above average (1172.5 msecs.) The difference between the pseudo-word latencies was also very small. Concerning this, Koda (2005: 185) asserts that pseudo-word naming is one of the most reliable measures differentiating strong and weak high school readers. Conversely, Siegel (1998: 146) reports that children with low scores on reading tests may not have poor recognition or decoding skills. Clearly these are very complex issues and surely cannot be reliably elucidated with 12-item sets of stimuli.

With regard to the two points above, I created my own version of the four lexical decision tests, preceded by a practice session. (Appendix 1) I included 20 items in each type of test. I ensured that all words had 4 letters, including the S/A task (unlike Shiotsu's which varied up to 9 letters.) All real words came from the top 2K bands of the JACET 8000 lists, and many from the 1K band (see below for a sample of my handout.) I imposed a uniform time limit to find out how many items could be completed in each category. I experimented until I settled on a 20 second time limit, in which only a small number of students were able to finish all the items in the real word decision condition, which was reported as being the easiest. Mean latency was found by dividing the number of seconds (20) by the score. The results were as follows.

Disregarding the slower latencies (probably due to time spent marking the page compared to tapping a key on a computer keyboard), what is arresting here is that scores for both the

Table 3: Score by Time limit (n. = 47)

	score (max =20)	mean latency (msecs)
Real Words	15.2	1316
Pseudo-words	17.7	1130
Irregular Strings	15.9	1257
S/A	11.0	1818

pseudo-words and irregular strings were higher than for the real words. This is at completely at odds with Shiotsu and Haynes' findings whereby the latencies for the first three types were progressively slower amongst EFL learners. I predicted my finding was due to a test habituation effect. Therefore I had another class of 24 students take the tests but with pseudo-words first, irregular string seconds and real words third. As I predicted, this time the score of the real word quiz was the highest, as is predicted by Haynes' original investigation using 48 items per set.

Table 4: Score by Time limit with the test order changed (n.=24)

	score (max =20)	mean latency (msecs)
Pseudo-word	16.4	1219
Irregular Strings	14.9	1342
Word	17.1	1169
S/A	11.6	1724

The differential between my two data sets is almost certainly due to test order. It is possible that in Haynes' sets of 48 items, also done on paper, the tests were long enough for the true relationships between the item types to appear. The difficulty for Shiotsu is that he used only sets of 12 items. Although his results mainly coincide with Haynes', I posit his findings are actually unreliable. These results also point out how the design of test instruments can unpredictably influence test data.

A more serious problem for Shiotsu's argument comes from Haynes and Carr (1990), an extended discussion of the research that is replicated here. Based on their results, they write, 'speeded tasks such as ··· lexical semantic matching of synonym-antonym pairs are related to the measures of reading speed, but not to the measure of *reading comprehension.*' (p.405, my emphasis added) Although Shiotsu provides apparently plausible results, it is reasonable to conclude that he was not measuring what he believed he was. Possibly his reading instrument was not as reliable as he thought, or that his synonym-antonym decision task, which came last in his battery, provided results which were compromised by the lack of test validity. At any rate, this research paper constitutes a warning about the dangers of extrapolating too many inferences from seemingly reliable instruments, when in reality multiple complicating issues may be present. Specifically, when attempting to measure word recognition efficiency, researchers need to construct tests that are as context-free as possible, since it is known that higher level processes in reading can obscure deficits in word recognition fluency. To achieve a 'clean' test (Stanovich, p487), it is necessary to account for memory, strategy, motivation effects etc. No test can be pure, but it is probable that Shiotsu's investigation became unexpectedly contaminated.

Critical Review 2: Jacobsen, 1995

In this short paper, Jacobson describes a fast and practical method of testing groups of people for dyslexia. The test instrument is comprised of two separate pencil-and-paper tests called Word Chain and Letter Chain. The relative performance on these tests is used to calculate a Word Recognition Index (WRI), with a low score indicating specific reading difficulties. Jacobson claims that the rapid completion time of the test (5 minutes) makes it ideal for classroom use in assessing recognition skill. The paper describes two experiments. One is a cross-sectional study of normal development of WRI, and the other explores WRI in dyslexic families.

Summary

To calculate WRI, subjects take the Word Chain test first. Each test item is a continuous string of three high frequency words (length 2 to 7 letters/word), such as:

boygomeet >> boy/go/meet

All words should be in the children's vocabulary. The subjects' task is to mark the boundary between the words with pencil strokes. There are 120 items and the test has a time limit of three minutes. Next, subjects take the Letter Chain test. Subjects have to segment a string of capitalized letters, marking the point where the same letter appears twice. For example:

OUCCNEMHHE >> OUC/CNEMH/HE

There are 80 items and the time limit is 90 seconds. A low score on the Word Chain and a normal score on the Letter Chain indicates word recognition problems. Jacobson claims that since the testing time in Word Chain (180 seconds) is 100% longer than the Letter Chain (90 seconds), a person with perfect word recognition ability theoretically should have 100% higher raw score in the Word chain. This would equate to a recognition index of 100. Individuals with the same score on both tests would get a score of 0. Young children or adults with dyslexia may have a negative WRI value. The formula for calculating WRI from the two component tests is:

$$WRI = 100 \times (WCh\text{-}LCh)/Lch$$

The first experiment focused on the normal development of performance on Word Chain and Letter Chain and WRI. Jacobson tested 150 school children ranging in age from 8-16 (grades 1-9). Data on college students, teacher students and teachers (up to age 65) were included. The results showed that ability on Word Chain progressed steadily from age 8 onward, peaking with student teachers in their 20s, before dropping back slightly in more mature people. Normal development on the Letter Chain Test was less marked and the change over time was much flatter, stabilizing after the age of 8 or 9. Girls did better than boys in each grade. WRI values rose rapidly from grade 1 to 5, and then were rather stable. In grade 2, the WRI was near zero, whereas in normal adult groups, WRI is about 90 indicating that word recognition is largely automatic.

The second experiment used the Word Chain Test to conduct an investigation of 32 American families comprising 140 individuals aged 16-75. 5 levels of reading and spelling skills were

identified. Level 0 had no problems, and were used as a control group. Besides the Word Chain Test, the test battery consisted of a spelling test, a non-word reading test and a test of intellectual ability. Those who failed in all three reading and spelling tests were classified as having the most severe dyslexia, belonging to level 4. Persons who succeeded with one of the tests were classified as belonging to level 3 and persons who succeeded in two of the three tests to level 2. Level 1 were people who self-reported as dyslexic but for whom disability was not indicated by the test battery. The results are shown in table 5.

Table 5. WRI and average scores of Word and Letter Chains for subjects classified on five levels of reading ability, aged 16-75

Level	WRI	Number of male/female	Average of Word chains	Average of Letter chains
0 (control)	83	24/40	64	35
1	66	5/7	65	40
2	60	17/11	56	35
3	34	15/18	42	32
4	4	11/2	32	30

An ANOVA showed significant differences between the 5 levels in WRI as well as Word and Letter Chains. The WRI is very low in level 4, corresponding to normal WRI for children aged about 8 years old. Jacobson reports that a WRI under 10 in adults seems to be a good indication of dyslexia, on condition that the number of word chains processed is below a certain limit. A low WRI combined with a low result on Word Chain has proved to be a quick and reliable method of identifying dyslexics. Many dyslexics also seem to have a slow processing speed in the Letter Chain Test. Low scores on both tests could indicate a generally low processing speed. It is noticeable from these results that the proportion of women to men markedly declines from level 0 to 4.

The reliability of the Word Chain was good. Test-retest correlation (Spearman) with a 12 month gap between measurements was r=.80 to .90 in different groups from grade 1-6. The figure for Letter Chain was somewhat lower. The correlation between the Word Chain Test and a 10 minute silent reading was r=.72.

Critique

In this section, I will outline some aspects of this test format which mark it out as very sensible. Next, I will highlight a problematic issue with the interpretation of the word recognition index. Finally, based on the results of a small-scale investigation of my students' WRI, I will ask whether this test could also be applicable to the measurement of EFL learners' development of word decoding ability.

First, this test format has several praiseworthy points. From a practical testing point of view, it is easy to dispense, easy to mark and calculate and is apparently highly reliable in its ability to distinguish different levels of severity of decoding disability. These features become relevant in light of the characteristics of dyslexia and its effects. People with dyslexia have difficulty in

decoding single words due to an insufficient knowledge of spelling-sound correspondences (Stanovich, 1982). However, people with dyslexia may compensate for this by using textual comprehension strategies, and go undiagnosed for a long period of time. The obvious advantage of a quick, easy test format such as Jacobson's is that it can be used to identify people with minimum fuss. If deployed in the classroom, teachers need little training in its use, and it can be marked fairly rapidly. Despite the large number of items on the test (120 and 80), with a little ingenuity marking of a large number of papers could be made very rapid with the use of a mask to place over each answer sheet to reveal if the segmentation marks are drawn in the correct place. Conceivably a single answer sheet could be done in 30 seconds, with the result that a class of 30 could be tested in a matter of minutes.

A related pertinent feature is the large number of words which appear in the Word Chain Test. In only 3 minutes, 360 basic items of vocabulary are potentially tested (120 chains × 3 constituent words). This ensures that a significant proportion of basic words that children or more mature people should be able to recognize are taken into account. The issue of the sample size in word tests is one which has been made by Meara (1996, p39-40). Where the sample is small (for example 20 items) this represents only 1% of the most frequently occurring 2,000 words. However, Jacobson's method would result in an impressive 18% coverage of the same range of words. In Jacobson's research on dyslexia, the size of the sample is an important factor. First, if an attempt is being made to diagnose the condition, it is important to base such judgments on as large a sample as possible. Second, many English words have pronunciations that are not rule-governed and the development towards automatic visual recognition is delayed in dyslexic people because of incomplete knowledge of spelling-sound correspondences. (Bruck, 1990, p440) A large coverage of words in the test would include a large number of such difficult-to-read items.

Notwithstanding the praise offered above, there is one aspect of the calculation of WRI which requires comment. Jacobson claims that a person with "perfect" word recognition ability should have a 100% higher raw score in Word Chain, resulting in a WRI of 100. Since the maximum score in Letter Chain is 80, a 100 higher raw score in Word Chain would be 160, 40 more than the number of items. A perfect score in both tests results, in fact, in a WRI of 50 as shown in hypothetical cases in Table 6. Further, due to variations in individuals (tiredness, interest, waning motivation etc.) it is conceivable two individuals could achieve the results in cases 2 or 3, resulting in a much higher or lower WRI. Also, it is difficult to know what to make of a WRI of 0 resulting from two different results on the Word Chain test, as shown in cases 4 and 5. Jacobson doesn't directly deal with this issue, only claiming that a 'WRI' under 10 in adults seems to be a

Table 6 Hypothetical scores and resulting WRI values

Cases	WCh score	LCh score	WRI value
1	120	80	50
2	75	45	66.6
3	70	55	27.3
4	70	70	0
5	45	45	0

good indication of dyslexia, on condition that the number of word chains processed is below a certain limit.' (p265) It is not clear where this limit might lie.

Although Jacobson presents this as a group test, the clear implication of his study is that this method can be used for identifying individuals with decoding inability. A further query is why Jacobson chose to use capital letters in Letter Chain.

I investigated these issues on a group of 35 non-English majors. I created my own version of the Word/Letter Chain test. (Appendix 2) The scores for WRI went from 51 to -44 with a mean of -10. There was a wide variety of score profiles among the students, and few had a predictable balance between Word Chain and Letter Chain. Some of the more perplexing results are shown in table 7. The student with the highest WRI (case 1) did so only on account of going very slowly on the Letter Chain section. The other two scored higher than her on Word Chain yet ended up with a lower WRI on account of correctly marking a higher number of Letter Chains. So it would appear that to assume a common speed factor between Word Chain and Letter Chain would not be valid among my Japanese students, at least.

Table 7: Real samples from English non-majors.

Cases	WCh score	LCh score	WRI value
1	53	35	51
2	63	63	0
3	57	76	-25

Although Jacobson's statistics revealed WRI to be reliable in the identification of already diagnosed dyslexics, I question if there is any causal relationship between performance on Word Chain and Letter Chain, even with his subjects. If there is, it does not seem to be present in the scores of the students presented above.

Finally, could WRI be used to track the development of non-dyslexic Japanese students, but who do have substantially incomplete word decoding ability? The same group of 35 students took QLex with 1st order approximation items. I found that there was a somewhat stronger correlation between Word Chain raw scores and QLex raw scores (Spearman Rank Order) $p=.59$ than between WRI and QLex raw scores, $p=.46$. This suggests that the Word Chain test alone measures a similar skill of word identification.

In conclusion, although Jacobson's format seems to be very praiseworthy there does seem to be some doubt about the inter-pretation of results, at least as far as applying the test to a Japanese context is concerned. This may be due to the complication of conflating two processes, word segmentation and letter segmentation, assuming they are comparable entities. At any rate, this issue reminds us of the complexities that are quickly encountered, in designing tests of basic recognition skills, when additional constructs are invoked by the design of the instrument.

Critical Review 3: Laufer & Nation, 2001.

The authors of this paper conducted a large-scale test to investigate the relationships between fluency and vocabulary size, and word frequency level. The aim of the test is to measure the

seed with which a subject matches a target word with its meaning. The instrument was a computerized version of the Vocabulary Levels Test ('VORST'). Overall, 'speed of retrieval' was moderately related to vocabulary size and word frequency. It also was found that the speed of NNS meaning recognition decreased with decreasing word frequency whereas that for NS was more homogenous across vocabulary frequencies. Content validity for the test is also claimed on the grounds that the recognition of word meanings, not only word forms, is tapped by this testing method.

Summary

Laufer and Nation (L&N) assert that it is important to be able to measure fluency of access as it affects language use, for example in reading comprehension. Word knowledge alone is not sufficient in this regard. They also claim their research is useful in addressing the theoretical issue of whether fluency is related to an integrated system of knowledge or whether it is item related. However, their interest is primarily in the practical issue of testing vocabulary performance in a way that approximates language use. Specifically, they aimed to develop a measure of 'word form recognition and the association of meaning to that form.' (p.10) L&N devised the Voca-bulary Recognition Speed Test (VORST), a computerized test incorporating a timing element. The basic procedure of the original VLT paper test is retained. However, along with the six word block, only one item is displayed on the left, unlike the three in the original format.

	1. apply
	2. elect
make_____	3. jump
	4. manufacture
	5. melt
	6. threaten

In response to this, the student should type in the correct number from amongst the choices on offer. Correctness of response is not shown. The program records the time from the appearance of each item to the moment a number (1-6) is pushed on the keyboard. The next screen is then displayed with the same block of six words, and a new word or phrase to match. After answering, a third screen is shown with a new item and the same block of six words. Once the three definitions have been matched, testees are offered the chance to amend their answers. If a new choice is made, this response time is substituted for the original time. The average response times in each block of six words and three items were recorded as were the average response time in the entire word frequency level. VORST is partly adaptive. In a given frequency level, if all 9 items are correctly answered, the program does not continue testing this level. However, if only one mistake is made, all six blocks are tested. This was to save time. L&N ask four questions in this research. These are:

1) Is there a difference in response times between groups with different vocabulary sizes?
2) What is the correlation between response times and vocabulary size?
3) What is the variance in response times of people with the same vocabulary correctness

scores?

4) Do the same people have different response times to words at different frequency levels?

Subjects with a wide variety of English vocabulary size were chosen. There were 454 subjects of whom 13 were native speakers. The rest were university students whose native languages were Hebrew, Russian and Arabic. Of these 12 were English majors, although the rest had passed a standardized entrance examination including English. None of the subjects were told that they would be timed although they were instructed to finish as fast as possible. The test has 5 levels (2K, 3K, UWL, 5K and 10K). Each level has 6 blocks with three items. This makes a total possible score of 90.

In the results, Kuder-Richardson 21 values for the correctness scores for the 5 levels of the test were as follows: 2000 level .82; 3000 level .80; UWL .73; 5000 level .79 and the 10,000 level .74.

Subjects were split into four groups: those scoring less than 15 were excluded. Group 1 had a score range of 15-32, group 2 33-50, group 3 51 to 68 and group 4 69-90. To investigate the first question on vocabulary size and speed of access, these 4 groups were compared on speed of response to words at the 3000 level and UWL. (The 2K level was excluded because it was found to include a familiarization time which skewed results.) The results are shown in table 8.

Table 8: Response times to words at the 3000 and UWL level

	Mean response time		Duncan Grouping			
Group	3000	/ UWL	3000 /	UWL	n	Vocabulary size score
1	18.44	23.09	A	A	188	15-32
2	16.63	22.995	A	A	86	33-50
3	14.45	20.50	B	A	89	51-68
4	8.19	12.97	C	B	55	69-90

At the 3,000 level the first two groups are not significantly different (as indicated by the Duncan grouping), although groups 3 and 4 are. The UWL level, which includes voca-bulary from the 4K and 5K levels, dis-tinguishes group 4, who have the highest size scores, clearly from the rest of the cohort. Fluency on the 3K level words increased when the average vocabulary size reached an average size score of 60. Based on these results, L&N claim that total vocabulary size has to be quite a distance beyond the tested level in order for there to be a significant difference in reaction time.

Concerning the second question of the correlation of response times and vocabulary size, the results in table 9 show a moderate correlation between knowledge at each level and speed of meaning recognition. However, there is a noticeably stronger correlation (-0.67) at the 10000 level although only 39 students were able to do this level. L&N claim from these results that people who know more words are more likely to access the meanings of words more quickly.

Concerning the third question of variance amongst people with similar scores, the authors selected NNSs who had at least 14 correct answers out of 18 at each level. Since very few were able to do the 10000 level, this was not considered. The 2000 level was also not considered because of the familiarization effect problem. As is presented in table 10, subjects who have the same score differ widely in the time they needed to recognize the meanings. In comparison, NSs

Table 9: Correlations between vocabulary size and speed of response in NNSs.

	2000	3000	5000	UWL	10000
r	-0.38	-0.40	-0.50	-0.31	-0.67
p	.0001	.0001	.0001	.0001	.0001
n	441	331	237	255	39

had very similar mean latencies of for each level of around 4.57 to 5.44 seconds. SD (1.4~1.6) and variance values (1.9~2.8) were also were far lower.

Table 10: Correct Response Time of NNSs

Level	n	answers	Latency Mean	SD	Variance
3000	54	18	12.2	6.4	40.8
5000	29	14	17.75	5.94	35.32
UWL	16	14	17.96	7.3	54.4

Concerning the last question of vocabulary frequency and response latencies, not all the testees were able to do all parts of the test. Further only those who had at least 6 correct answers at a level were compared. Table 11 shows the mean differences in response times between each pair of frequencies. The significance of the difference in indicated by **($p<.01$), *** ($p<.001$). A Repeated Measures comparison of four levels (3K, 5K, UWL, 10K) was performed on 35 subjects. There was a significance difference between the four response speed means $F(32,3)=23.09$, $p<.001$. All means except UWL and 5K were significantly different from each other. To compare more learners, 178 were compared on 3K, UWL and 5K. This too produced a significant difference ($F(176,2)=61.69$, $p<.001$). 225 students could only do the 3K and UWL levels and they also produced a significant difference with a t-test value ($t=12.04$, $p<.001$). These results show that the less frequent words take longer to match with their meanings.

Table 11: Differences in mean latencies between pairs of frequency levels in NNSs

Frequency level	Difference		
	A	B	C
	n=35	n=178	n=225
UWL>3000	5.15**	6.28***	6.6***
5000>3000	3.57**	3.84***	
UWL>5000	1.58	2.43***	
10000>3000	9.84***		
10000>5000	6.27***		
10000>UWL	4.68***		

Critique

In this section, I will consider VORST from the two perspectives of content and construct validity. Although this paper provides some impressive evidence in answer to the four questions asked, I have some strong doubts about the use of the Vocabulary Levels Test format for the purpose of measuring speed of meaning recognition. Not least of these is that research on the format of the traditional VLT is still ongoing and is revealing important limitations about its capability as a test instrument merely of vocabulary size. The VLT was originally intended as a simple diagnostic tool for classroom teachers, not as a sophisticated tool for addressing complex psycholinguistic issues. Nevertheless, L&N blithely press the VLT into use as a test of the complex relationship between semantic access efficiency and vocabulary size. This is something the format is specifically not designed for. This does not a priori render the findings of this paper invalid, but we must consider this issue quite carefully.

Certainly this incarnation of the Vocabulary Levels Test is another interesting, and idiosyncratic, contribution from Laufer and Nation. In the past, they adapted the VLT to the measurement of productive knowledge (Laufer and Nation, 1999). In that case, the claims made for the test received unenthu-siastic reviews (e.g. Read, p.125). This was partly due to the odd task whereby testees had to complete partially deleted words. The task was inconsistent because the blanks to be filled in were not of consistent length, resulting on confusion about whether the test was actually one of productive knowledge. Although learners at higher levels of proficiency did better than those of lower proficiency, in the words of one prominent commentator, 'this finding does not give any specific insight into the meaning of the test scores.' (Read, 2000, p.125)

Similar doubts can be expressed for VORST. Although the main finding of this research (speed on a frequency level increased only when learners' vocabulary size progressed far beyond that level) is very interesting, there is reason to be cautious. For example, the demands on the testees are not consistent through the test. The length of the six definitions in each block varies, with the highest frequency 2K type having an average of only 2.7 words for each definition while the 5K has 4.6. (figures taken from Nation's original 1990 version.) This could make a considerable difference in reading time, especially if testees read each definition twice. A related issue is that L&N claim they are interested in 'a measure that closely approximates language use.' (p.10) By this, it seems they are trying to overcome the issue of lexical vs. semantic access in word recognition testing by having testees match words to their meanings. However, when the definitions vary so much (from 1 word to as many as 9), the central issue of how much time is spent retrieving words from memory is obscured by cognitive processing effects entailed in comprehending the definitions. Another confounding issue is that testees are offered a chance to amend their answers at the end of each block with the new latency for that item replacing the original value. Finally, L&N make a surprising amendment to the original format whereby 9 straight correct answers (from the 18 items per level) are taken as evidence of mastery of the level. Testees are then relieved of the requirement to finish that block in order to 'save time'. The commonsensical approach to vocabulary assessment is to sample as many words of a learner's knowledge in as efficient a manner as possible. Here L&N take the opposite tack. The original form of the VLT samples 18 items per level plus the vocabulary needed to understand the definitions. This is only about a 2% sample if there are 1,000 words in each level. Despite this low figure, they settle for 9 words if the learner happens to answer the first 9 items correctly.

All told, VORST does not conform to Stanovich's call for 'clean tests'. In the business of measure very specific psychological phenomena, we need to keep our measures as simplified as possible. At least on this front, VORST lacks transparency.

If we now consider the test from the perspective of construct validity there are more reasons to be cautious. As we have seen, L&N highlight the two variables of word frequency and recognition time, and imply that the interaction of these two factors alone is responsible for the results presented in the paper. A key variable left completely unmentioned in this paper is guessing. Research on answering behaviour in VLT has shown that item dependence in clusters can result in testees having as much as a 50% chance of guessing an item correctly, if they have confidently answered the first two items. (Kamimoto, 2006)

Further, Schmitt et al. (2001) indicated that high proficiency students might have a greater propensity to guess successfully. I tried to investigate how the 10K level might be answered by interviewing an acquaintance with extensive overseas study experience in English literature. I rewrote all the items from Schmitt et al's 10K level (they had 10 clusters with 60 items) in a long, unbroken list and showed them to her. I asked her to mark each item with a '✓', 'X' or '?' to indicate her degree of knowledge of these items. This produced the following results:

Table 12: Results of a high-ability Japanese subject

	✓	X	?	
total	29	17	14	60
comments	7 words misread as other words		No correct translations of any item offered.	

After she did the 10K level, and despite being asked not to guess, she finished all the test items having successfully guessed 13 items, yet her overall score was only 20 out of a possible 30. However, she was still unable to provide any translation for most of these guessed items in Japanese. What I understood from her was that her advanced morphological knowledge of English allowed her to creatively navigate her way around the test, selecting words that seemed to match, but in many cases without having any idea of their meaning. For example, 'benevolence' was correctly matched based on a supposed association to 'benefit', 'salve' on an association to 'salvation' and 'vindictive' on an association to 'victim'. This is clearly very problematic. The greater general sophistication of morphological knowledge that accompanies the ability needed to respond to items at this level appears to allow some testees to initially bypass the need for any partial knowledge (the construct VLT wants to examine) and instead encourages answering on the basis of tangential hunches, which nevertheless seem plausible to the subject. I would argue this condition does not fall within 'partial' knowledge. In the case of my informant, the construction of the test has apparently distorted her knowledge into a higher score, with the construction characteristics of the cluster having induced her to answer, which I posit, is different from guessing.

What this analysis shows is that, in addition, to guessing, there are other strategies which testees may use through the lower level frequency items with unknown effects on speed of answering. If even partial knowledge of items cannot be vouchsafed by this method, it makes no sense to attempt to link word meanings to recognition speed, and further to draw conclusions

from the comparison to high frequency levels concerning the development of the L2 lexicon. Such a mutually exclusive link simply does not exist.

In conclusion, VORST provides tantalizing views into the relationship between size and access speed to words in memory. Describing the contours of such a relationship is of too crucial importance to take these findings on face value, even though the results do point in an intuitively appealing direction. The reason for this is that as an instrument, VORST is almost certainly affected by confounding factors that the authors simply pay no attention to.

Interlude

One of the problems with studies such as those by Laufer & Nation and Shiotsu is that their results are compromised by controllable factors to the extent that it becomes unclear what exactly is being measured. In this sense, their tests lack validity by failing to strictly follow Stanovich's dictum of a 'clean test'. The interactionalist position of researchers such as Chapelle (1998), who insist that vocabulary ability must only be investigated in the context of real reading tasks, creates difficulties that unnecessarily obfuscate specific processes of L2 lexical development. Chapelle claims, for instance, that lexical recognition can only be considered while a subject is actually reading since mental processes will differ qualitatively depending on context and demands of the task. My answer is that this view assumes that reading is a primarily top-down process, and moreover contextual factors such as topic and content difficulty, and individual factors such as background knowledge or motivation exert primary influence on reading and comprehension. Yet, this view disregards the wealth of research findings which demonstrate that word recognition skill alone does not have a strong causal relationship with reading comprehension. Word recognition skill is but one of many important elements that make up the suite of necessary skills for reading. It is quite feasible that a good reader can have word recognition speed that is markedly slower than the average for people similarly matched in reading comprehension skill, and wider L2 proficiency. As learners become more advanced these individual differences undoubtedly subside. Yet, at lower levels these differences can go unnoticed and possibly untreated.

One important purpose of L2 word recognition research is to elucidate the cross-linguistic effects on reading between different orthographical systems. (e.g. between Japanese and English) The native language of a L2 learner 'embeds habits of mind, instilling specific processing mechanisms' (Koda, p9) that remain engaged even while reading in the new L2. The implication is that proficiency in L2 decoding does not appear as a direct causal effect of experience in the new language: language processing and linguistic knowledge are separate competencies. Further, people who are proficient at reading in their first language will not necessarily be good in their second. Another key issue is the importance of phonological awareness of the new L2 orthography, and its connection to rapid word recognition ability. Those students for whom phonological decoding is influenced by their L1 may, ironically, end up over-relying on context as a compensatory strategy (Koda, 2005). Although use of context to aid comprehension is regularly recommended for readers at any level (e.g. figuring out the meaning of an unknown word or for 'reading between the lines'), this popular advice is strongly rooted in L1 reading research where decoding issues are not nearly as pressing. The advice of Goodman (1967) that reading is a 'psycholinguistic guessing' game, in which readers try to predict meaning based on global text features, in the way that learners do in L1, de-emphasizes the

importance of the bottom-up aspect of text-decoding for foreign language learners. Even the reading processes of advanced learners of English are influenced by their L1 habits. At lower levels, inefficient reading places a heavy burden on short-term memory, making comprehension difficult. This results in large individual differences the reasons for which are not always clear to teachers, learners or indeed researchers. Chapelle makes no concession to this consideration. Yet it would seem far easier to define constructs by carefully attempting to isolate them in tests designed to be as independent of context and content as possible, so that investigations are not hopelessly clouded by extraneous factors.

Part 2
A test of lexical accessibility through ability of visual word-recognition

Introduction and Background

This section reports the results of a timed word recognition test, 'Q-Lex'. In this research, accessibility of words is taken to be one of the three components of lexicons, along with size and organization. (Meara, 1996) The focus here is not with the reaction time of individual words per se, but with average reaction-time across all test items. The greater number of words that a student can identify within the time limit, the more automatic recognition is.

In lexical decision tasks, the difference between native speakers and L2 learners in reaction time is in the order of tens of milliseconds. This is too small to be accurately gauged without specialized equipment. Therefore, even if speed of accessibility of words improves with increasing proficiency, and thus potentially is a relevant measure of progress, tracking the development of this aspect of learners' lexical competence has not been possible for teachers. Performance in an earlier, similar test of accessibility was found to correlate with other tests of 2nd language performance. (Lambert, p203, 1990)

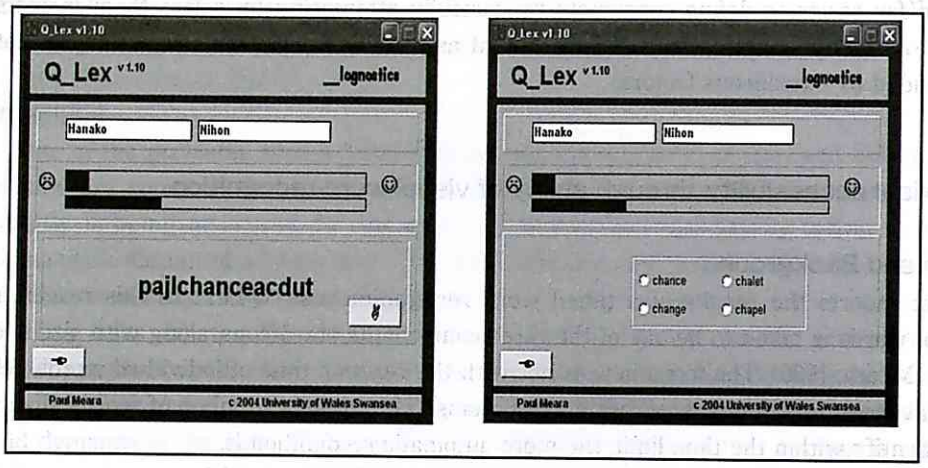
Q-Lex is designed to measure speed of recognition using a personal computer, making it suitable for tracking the progress of students enrolled in language courses. The test presents 50 high-frequency words. These words are hidden against a mask of surrounding letters, as in this example: pajlchanceacdut where the hidden item is 'chance'. As quickly as possible, testees click with a mouse to stop the timer and select the correct answer from a choice of four words, as shown below. The masks are designed to delay recognition-time to a degree measurable by a personal computer.

In the test, all words are tested on native speakers. Especially for basic vocabulary, the recognition time of native speakers represents an objective standard against which it can be measured how automatized students' recognition ability has become. The average reaction time from native speakers is taken and from this a norm value for each word is calculated using the formula ((2 × SD)+RT). As an example, a word with a mean native speaker reaction time of 1000 milliseconds may have a standard deviation of 250. In this case, the norm for this item would be 1500 milliseconds. If a student's reaction time in the test to the same item is less than 1500 milliseconds, she is credited with 1 point. If it is more than 1500 milliseconds, she gets 0. (2 standard deviations around the value is taken as being statistically related to a given value, by convention.)

Q-Lex makes it possible to investigate patterns of change in accessibility over time, for example whether access initially improves gradually or rapidly. If the test is to prove useful, it is

important to investigate the how it performs in various situations. This experiment will focus in particular on consistency of recognition of the same items over time. The failure to re-identify words in subsequent tests is problematic since consistency probably implies answering the same items in repeated tests. This is turn may indicate stability of knowledge of words.

Fig 2.1: Q-Lex test screens showing the search and multiple choice screens.



Early Results and Challenges in this Research

The most important aim of this research is to establish that Q_Lex can reliably measure progress in word recognition over time. This has been provisionally established by the findings shown in table 2.1. (Coulson, 2005) Two classes of English-major students, one first-year, the other second-year took the same test. The abilities of the second-year class were similar to the first-year class based on the results of a well-known proficiency test taken when they were in the first year of their course. The investigation revealed that a year of full-time study resulted in a score of 22.8 for the second-year students compared to the first. Further reaction-time latencies also showed a difference based on an extra year of study.

Table 2.1. Mean number of items recognized by students taking the test for the first time. (n.=93)

	2004 year group (n.=50)	DC05 year group (n.=43)
Mean score (max. 50)	17.2 *7.9*	22.8 *9.8*
Mean Reaction Time (MS)	2070 *348*	1930 *341*

However, the subjects' scores on Q_Lex often lack consistency. That is, the second time subjects take the test, to varying degrees they fail to answer the same items within the statistical norm that they answered them the first time. This is an important issue since I would expect words that are automatically accessible to students should be recognized much more consistently than they are. Certainly, randomly recognizing items in different tests does not instill confidence that this test reliably taps the ability of students to recognize the words in masks. There may be several factors behind this. Weaker students whose automaticity in English word-recognition skill is less developed may have trouble in accessing their L2 lexicon

as reliably as more advanced learners. A more troubling cause of this phenomenon may be the design of the test itself. Therefore, it is important to elucidate this issue.

In earlier experiments many students reported being able to identify the English-like letter combinations of the beginning of the target word buried in the string. This strategy of correctly answering the test items falls short of what is required in a test of (whole-word) word recognition. The letters of the surrounding mask frequently do not combine with the letters of the hidden word to create highly probable letter combinations in English. For example, the word 'family' which is contained in the following item utrfamilypoghea is particularly conspicuous due to the incongruous surrounding letters of the mask. The first plausible English-like combination to appear is 'fam'. On the basis of only this, it appears that some subjects have been pushing the 'stop' button.

There are two ways around this. The first one is to devise entirely new masks for the test words which blend more effectively with the target word. The aim here is to blend the word more naturally with the target word.

The other method is to use distractors in the answer screen which, as far as possible, share the same first two or three letters of the target word. For example, for 'arrive' the four multiple choice answers are [arrive.around.artist.arrest]. It is hoped that this will discourage testees from attempting to rely on syllable recognition when they are confronted with various alternatives that begin with a similar letter combination.

The Methodology and Development of Masking Strings

The masking strings which have been used so far in this research are composed of first-order approximation ('1OA') letter strings. Here, the letters which surround the target word have been selected at random from an English text. This is different from simply picking letters at random since an English text naturally reflects the statistical distribution of letters in English. Miller (1963:85) The letter 'e', for example, will be far more commonly picked than a letter 'z'. This results in strings which do not closely resemble English-letter combinations. As in the example discussed, above the letters surrounding 'family' do not closely resemble plausible English letter combinations.

It is also possible to make second-order approximations are made in the following way. From an English text, a double-letter combination is chosen. For example, 'se' is chosen (from 'Second' at the beginning of this paragraph.) The next letter must be one which is statistically likely to follow the letter 'e'. The text is scanned to locate the next example of a letter 'e', and the following letter to it in that word is selected. This gives 'ser', (the 'r' being taken from the word the word 'order'.) The next occurring 'r' appears in 'approximation' and the following letter 'o' is appended, making 'sero'.

This method quite frequently results in real 3- and 4-letter words. These can appear as independent words in the surrounding string or can be extensions of the hidden 6-letter word. In either case, they are an obvious distraction from the task of finding the intended word, and need to be removed to leave a string which has only the appearance of English non-words. In the item ithattackievesp 'attack' may be most obviously visible but the presence of 'th' before 'attack' combines to create a distracting 4-letter word 'that', which may attract the eyes of EFL students more than native speakers, particularly if they scan left to right rather than rely on whole-word recognition.

The ideal item is one in which the target word is masked by letter combinations which blend more authentically with its border letters. It is hoped that the testee may need to identify the target word more completely, rather than rely on constituent syllables. The word 'family' in a 2OA string becomes: lolfamilypededede This should be harder to identify than its counterpart 1OA item utrfamilypoghea discussed earlier.

The masking methodology is not, at any rate, perfect since in both first-order and second-order approximation strings the juxtaposition of certain letters at the word and mask border, such as consonants or vowels, may have an effect in making the word more or less obvious. However, it is difficult to predict which items will better discriminate learners. Therefore for this experiment, no further selection of items was attempted for the test, once the items had been constructed. Concerning the masking methodology, the hypothesis tested in this experiment is that 2OA will be answered more slowly, but more reliably than 1OA strings.

Method

50 words for testing were selected from the top 1K and 2K bands of the JACET 8000 (2005) vocabulary frequency list. The mean frequency of these words was 950. These words were embedded in both 1OA and 2OA strings, to create two parallel sets of items. They are shown in appendix 3. The items were all six-letter words.

45 students took the version with first-order approximations strings and 43 took the test containing the newly made 2OA strings. Both groups were from the same year and were equivalent in terms of TOEIC scores; (458 vs. 467). The second 1OA test was administered 6 months after the first to 35 of the original students. The second 2OA test was administered 4 months after the first to 19 of the original students.

Data from 60 native speakers mostly in their 20s or 30s was gathered. 29 took the 1OA strings version, and 31 the 2OA version. Reaction time norms for each of the 50 items were calculated by taking the mean reaction time of native speakers and adding to it twice the standard deviation of the reaction time. ((2 × SD)+RT). Outlier values in native speaker times were removed by the Smirnov-Grubbs test at the 0.01 significance level. With this test, values that *look* extreme and certainties for rejection sometimes prove not to be. Conversely, the Smirnov-Grubbs test sometimes identified one or more outliers in the same set of one native speaker that looked quite reliable. Overall, the mean time of the 50 norm values increased from 3.0 seconds to 3.8 seconds in 1OA items, and from 3.4 seconds to 4.6 seconds for the 2OA items, after the removal of outliers by this method. The largest norm value in the 1OA set was 9500 msecs and in the 2OA set 16200 msecs (16.2 seconds). See the discussion section for more on this issue.

The masking strings were made by sampling the text from a popular science book. The target words were then inserted at different positions in the strings to ensure students were not able simply to identify the words by looking at the same place each time. The position of the same word in the 1OA and 2OA strings was the same to make comparison easier. The three distractors for the answer screen were made as similar as possible. They were also 6-letter words, and as far as possible started with the same syllable. However, sometimes it was impossible to find three 6-letter words which all began with the same three letters. For example, the first item in the test is *change*. Here the distractors are *chance*, *chatty* and *chunky*. On occasion it was necessary to choose distractors which students were unlikely to know since no other alternatives could be found. For example, in the set of alternatives for the item 'dinner'

[dimmer.dining.dinner.dinghy] I thought it unlikely that either 'dimmer' or 'dinghy' would be known by any subject. So effectively, the only likely choices for confused testees are 'dining' (in fact 7 letters) or the correct answer 'dinner'. Where possible I constructed more satisfactory alternatives than those shown here, although many initial 3-letter combinations only have low frequency vocabulary items. (See appendix 4 for a full list of items and answers.)

Results

This section will start by looking at the raw scores of students on both kinds of tests. Next, it will present the results of consistency of recognition hits across test pairs. This will include a focus on the performance of individual students on both kinds of tests. Finally, the relationship between test scores and mean reaction times on the tests will be considered.

1) The results for the 1OA and 2OA items are shown in table 2.2.

Table 2.2 Scores for the two tests

	1OA strings		2OA strings	
	Mean score	Highest/lowest score	Mean scores	Highest/lowest score
Test1 score	20.7 *8.2*	43 / 7	18.3 *7.8*	38 / 6
Test2 score	26.9 *8.3*	46 / 11	27.8 *9.2*	43 / 14

The results in the first test showed that the less English-like strings of the 1OA items allowed students to recognize 2.4 more items (a 13% difference) than 2OA did. However, there was a notable difference in the second test where students in the 2OA test found 0.9 more items on average. The reliability (kr-21) for both tests was good. (1OA=83.3; 2OA=82.5) Almost all students got a higher score on test 2 than test 1. Only one student's score on the 1OA set fell from 31 to 20 points.

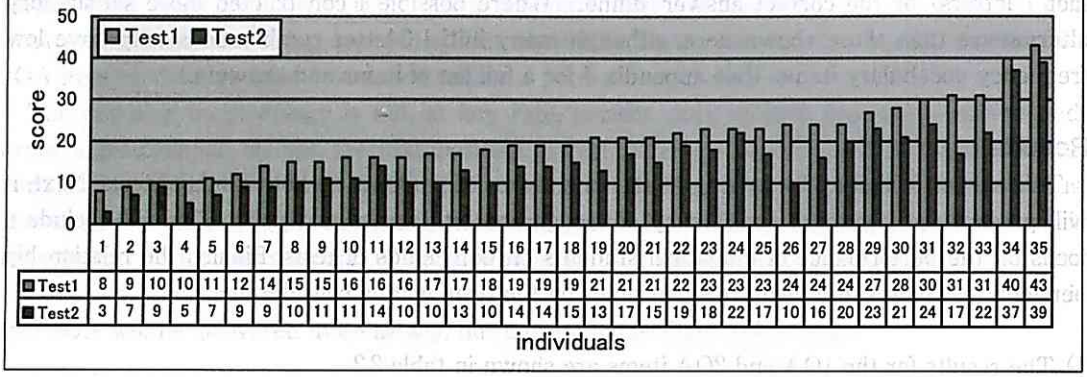
2) The issue of performance consistency was investigated by checking the number of items which students recognized in both tests. The results for both types of strings are shown in table 2.3.

Table 2.3 Consistency in item recognition

	1OA strings	as a percentage score in test 1	2OA strings	as a percentage score in test 1
score	15.1	73.1%	12.6	70.8%

As figures 2.2a and 2.2b show, variation between students, in their ability to find items previously recognized, was quite marked in both 1OA and 2OA sets of items. Students who performed best on the 1OA (figure 1a) test with an initially high score in test 1 (numbers 34 and 35 who scored 40 and 43 points) found, unsurprisingly, a high proportion (91%) of the same items again. Others who had a lower score, such as numbers 2, 3, 11, 22, 24, 28 & 29 still showed high consistency (87%) in answering the same items again. Conversely, others such as 1, 15 and 32, showed quite poor consistency (52%) in re-answering items recognized earlier.

Figure 2.2a Number of items recognized again over 2 tests in 10A strings

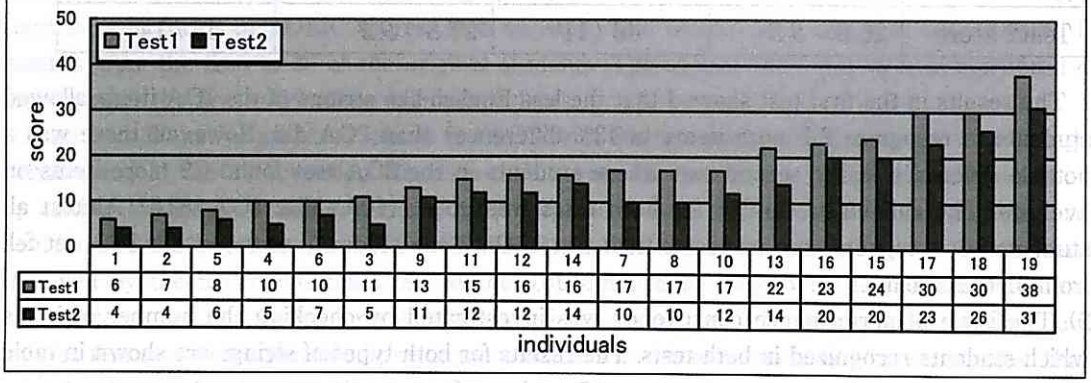


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Test1	8	9	10	10	11	12	14	15	15	16	16	16	17	17	18	19	19	19	21	21	21	22	23	23	23	24	24	24	27	28	30	31	31	40	43
Test2	3	7	9	5	7	9	9	10	11	11	14	10	10	13	10	14	14	15	13	17	15	19	18	22	17	10	16	20	23	21	24	17	22	37	39

individuals

In the graph for the 20A items (figure 2.2b) there were also some individuals (1, 9, 11, 14, 15, 16 & 18) who were particularly capable, answering a high proportion (84%) of the items they had recognized before. Conversely, other students (3, 4, 13) recognized only an average 56%.

Figure 2.2b Number of items recognized again over 2 tests in 20A strings



	1	2	5	4	6	3	9	11	12	14	7	8	10	13	16	15	17	18	19
Test1	6	7	8	10	10	11	13	15	16	16	17	17	17	22	23	24	30	30	38
Test2	4	4	6	5	7	5	11	12	12	14	10	10	12	14	20	20	23	26	31

individuals

Concerning the issue of reliability, the tables below present the performance of three individuals across each of the two tests: the highest scoring, median and lowest scoring students. They show the number of words which are:

1) recognized at T1
2) recognized at both T1 and T2
3) recognized at T1 but not at T2, or recognized at T2 but not T1
4) not recognized at T1 or T2

Comparing tables 2.4a and 2.4b, we see that the 10A items are generally easier to recognize, and are more reliably recognized in a second test. This is especially so with the strongest and the middling student. Weak students perform poorly in both tests, but it is noticeable that in the 20A set, the student had lower rate of recognizing items for the first time in test 2 than her

counterpart in 10A (129% vs. 238%). This reflected in the fact that 20A test has a higher rate of items going completely unrecognized in either test, at least amongst the weaker students.

Table 2.4a: The performance of single subjects in 2 tests with the 10A set.

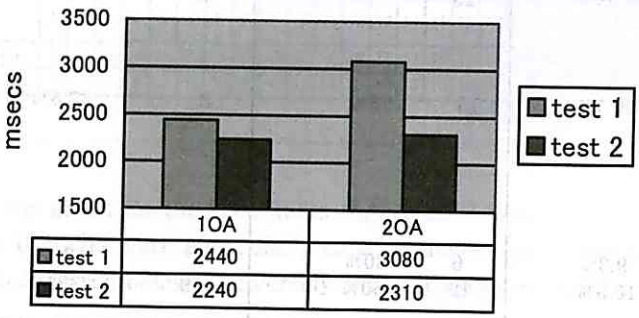
	The highest scoring student		A student who scored in the mid-range		A low scoring student	
	Score	% of T1score	Score	% of T1score	Score	% of T1 score
1) T1	43	----	20	----	8	----
2) answered in T1 & T2	39	90.7%	15	75%	3	37.5%
3) answered in either						
T1	4	9.3%	6	30%	5	62.5%
T2	7	16.3%	12	60%	19	238%
4) not recognized at T1 T2	0		16		23	

Table 2.4b: The performance of single subjects in 2 tests with the 20A set.

	The highest scoring student		A student who scored in the mid-range		A low scoring student	
	Score	% of T1score	Score	% of T1score	Score	% of T1 score
1) T1	38	----	17	----	6	----
2) answered in T1 & T2	31	81.6%	12	70.6%	4	66%
3) answered in either						
T1	7	18.4%	5	29.4%	3	42.9%
T2	12	31.6%	11	64.7%	9	129%
4) not recognized at T1 or T2	0		22		34	

Consistency in performance can also be assessed by looking at the speed with which students recognize the items. Figure 2.3 shows the mean latencies of students on items recognized in T1 and T2. Clearly, there is significantly more difficulty in recognizing the 2OA items than the 1OA items at first, yet this effect largely disappeared in the results for the second test where there is a differential of only 70 msecs between 10A and 2OA mean latencies.

Figure 2.3: Mean RT of items recognized in both tests



	1OA	2OA
test 1	2440	3080
test 2	2240	2310

3) The third set of results concerns the relationship between test scores and mean reactions times. First, figure 2.4a shows students' mean reaction times against NS mean reaction times and their associated norm values for each item. (The last 2 norms are off the scale.) The first 16 mean values for students have a similar degree of facility, with very similar mean recognition times. The pattern of these items' latencies are close to that of native speakers', at around 1500 msecs. From the 22nd item onward, the pattern of students' mean reaction times becomes increasingly unstable, as does that of the native speakers after item 33. Nevertheless, almost all the mean reaction time values of students fall well within the norm value. There are only two noticeable exceptions (17 and 28), but they too fall just within the norm limit. Somewhat surprisingly, the final item shows the mean student time is faster than the mean native time. This is due to the fact that one native informant in particular took exceptionally long to identify one item (although she was not an outlier.) This also explains why the norm for this item (9549msecs) is off the chart.

The parallel results for the 20A string set of items (figure 2･4b) are not dissimilar although here it is noticeable that students' mean reaction times are not as stable in the early stages. However, there are two instances, items 10 and 12, where the students' mean time is better than the natives'. In both cases, 7 students answered the item. With both items several students recognized the items exceptionally quickly, at between 700 and 900 milliseconds

Figure 2.5 shows the distribution of successful recognition hits by three students on the 2OA set of items. These are the same students who appeared in table 2.4b. The hits of the most competent student (score 38/50) are spread fairly evenly across the board. Moreover, most of her hits actually fall below the mean reaction time of native speakers. In contrast, most of the hits of the average performing student (17/50) fall in the right-hand side of the graph, and are almost all (14) above the line of native speakers' mean reaction time. Finally, the weakest student's recognition hits are also on the right-hand side, but are too few to discern a clear

Figure 2.4a: Comparison of students' mean RTs against NS mean RTs and associated norms: 10A strings

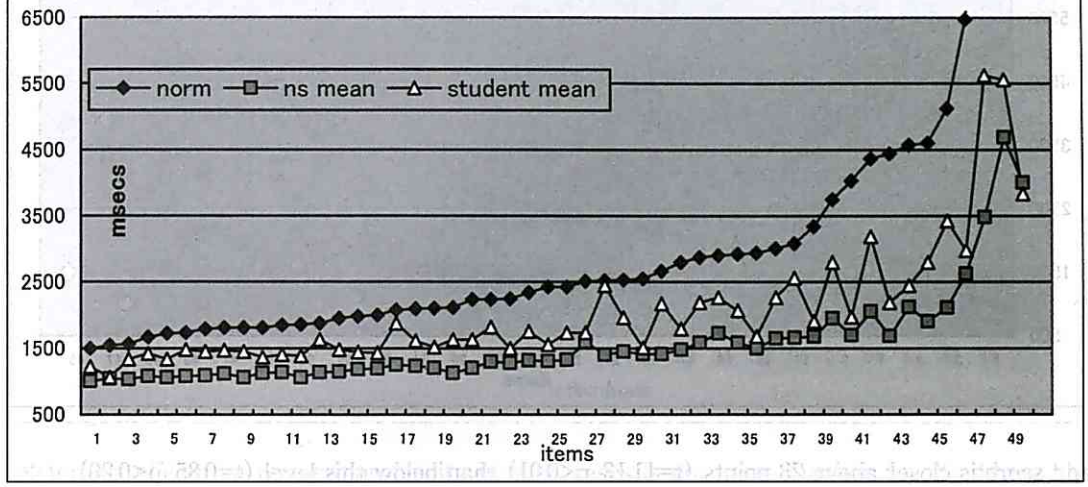
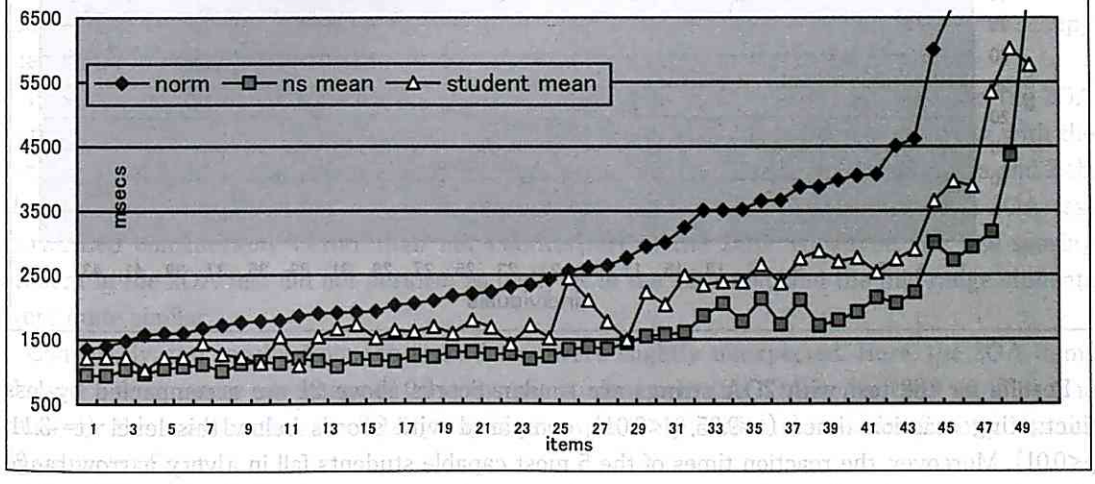
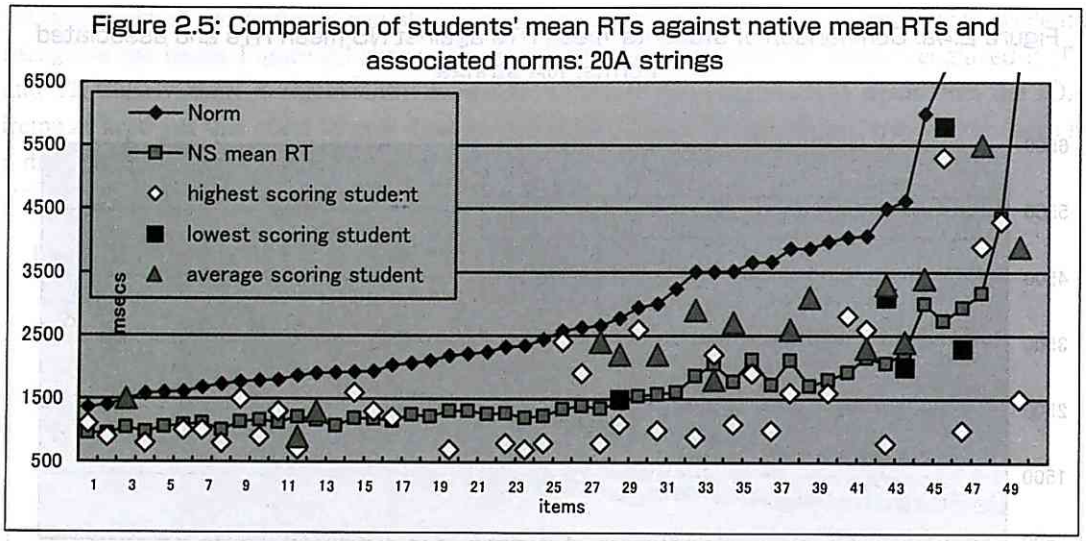


Figure 2.4b: Comparison of students' mean RTs against NS mean RTs and associated norms: 20A strings

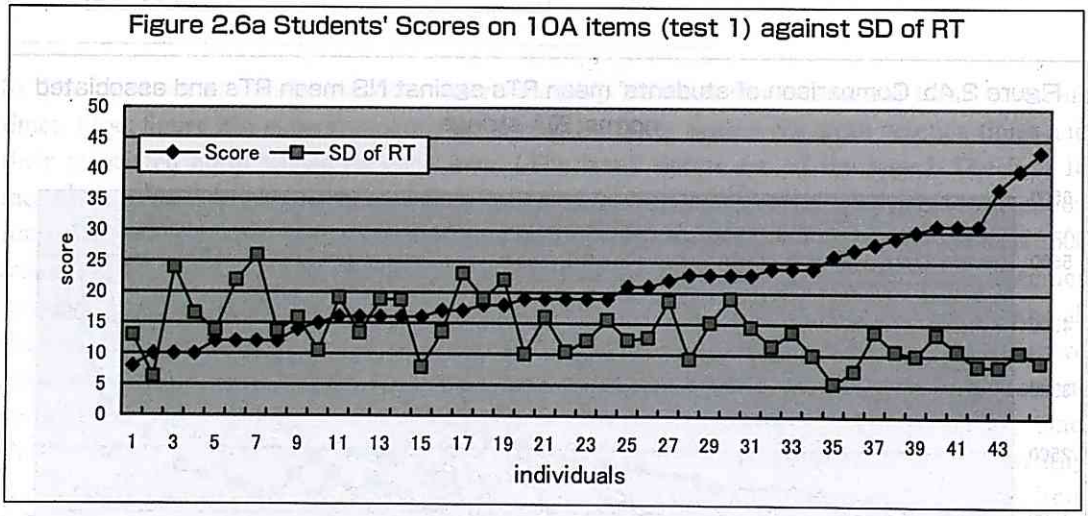


pattern. The parallel analysis for the 10A test produced largely similar results with the difference that the hits for the average scoring student were spread more evenly across the items from left to right.

Scores above a certain level on Q_Lex are accompanied by more stable mean reaction times than lower scores. Figures figure 2·6a and figure 2·6b show the scores of students plotted against the standard deviation of reaction time for all 50 items on the test. The graph for the 10A items test (figure 5a) shows that scores of above 23 points are accompanied by a much lower degree of variation in reaction time than for lower scores. The relationship between SD of reaction time



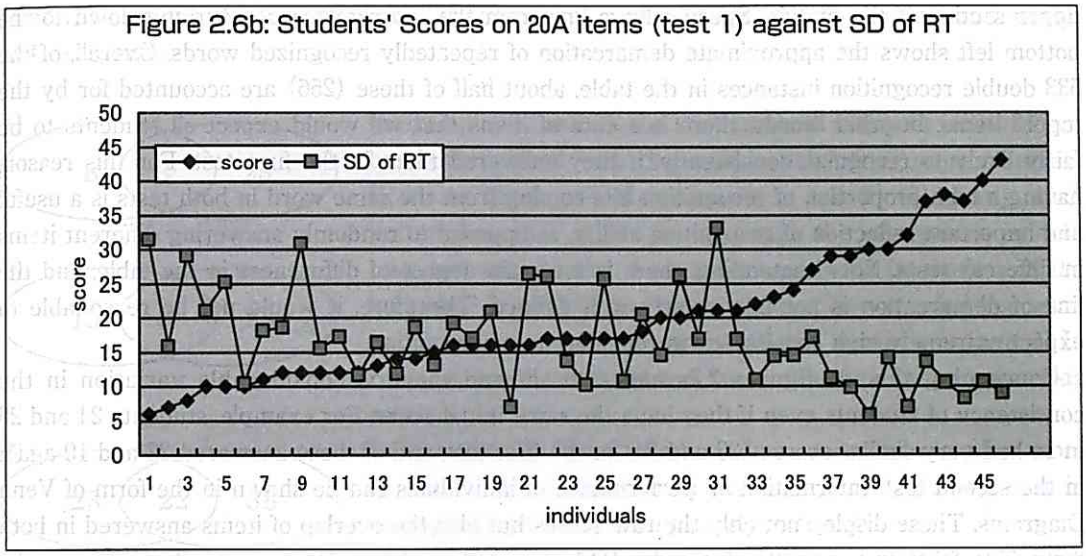
and score is closer above 23 points ($t=11.13$, $p<0.01$) than below this level ($t=0.85$, $p<0.20$).



Results for the test with 20A strings are similar. Scores above 21 are accompanied by less fluctuating reaction times ($t=9.05$, $p<0.01$), compared with scores below this level ($t=-3.11$, $p<0.01$). Moreover, the reaction times of the 5 most capable students fall in a very narrow range of consistent reaction times.

Discussion

In this section, I will discuss the rather complex picture that these findings present. Three points, in particular, will be taken up below. First, although the 20A strings were intended to increase the accuracy and consistency of item recognition, the 10A strings proved more reliable overall, although clarifying why is important. Second, looking forward to the next round of experiments, we can ask which items perform best, and why. Clearly some items do not perform well. However, identifying a systematic pattern for why particular masking strings have



deleterious effects on recognition facility in such cases is not easy. Further, I will address the issue of identifying the best performing items so far with the aim of making a new 'best-of' reliable set from those which have already been tested. To do this, relying solely on the number of recognition hits each item receives may not be a sufficient criterion. Therefore, I will make use of Rasch analysis as an additional perspective. Third, we need to consider how to interpret the performance of all individuals in the test group. Although the results at the extremes of the score range are very clear cut, there is a need to address the majority of students who occupy the middle ranges.

First, in test 1, the 10A items produced 2.4 extra items, or 13% more, compared to the 20A variety. (20.7 vs. 18.3) This is was predictable since 10A strings combine less naturally with the target words making the targets stand out more from their back-grounds. In tables 2.4a and 2.4b, that showed the performance of single subjects, the highest scoring student in the 10A test performed considerably better than her counterpart in the 20A. However, the low scoring student in the 20A test did not perform as badly as in the 10A test, and the mid-range students were quite similar.

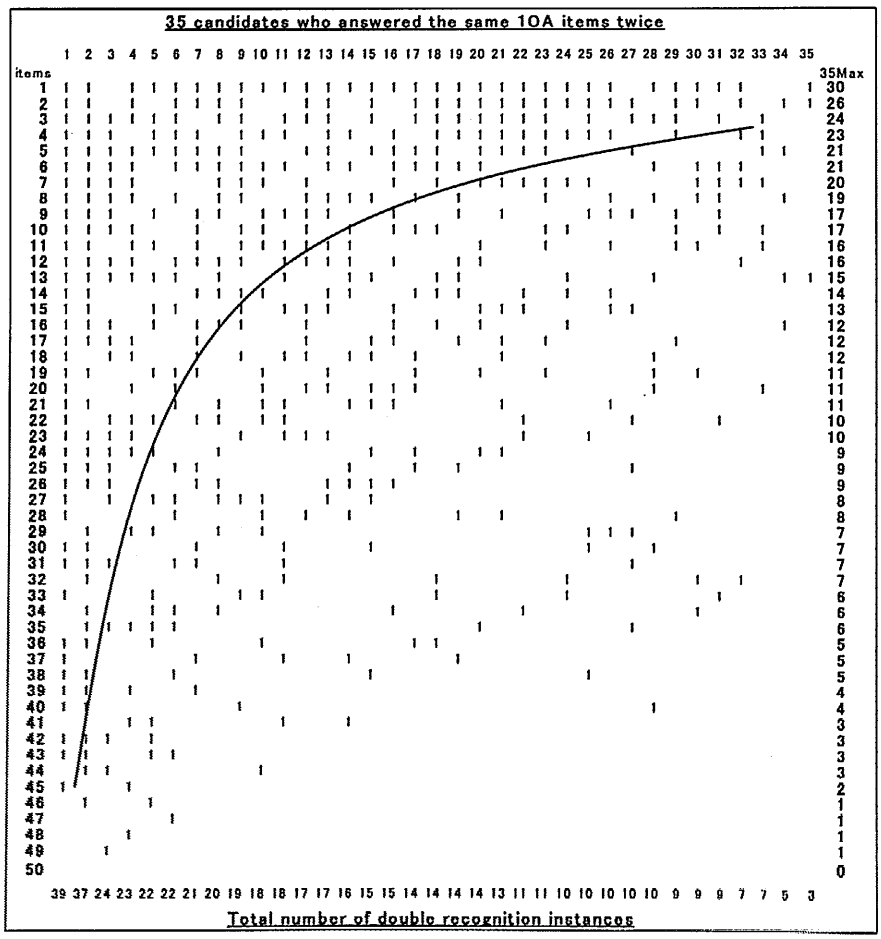
Conversely, the results from the second test were slightly unexpected. Here, the 20A items showed a mean score advantage of 0.9 items per student over the 10A items. (26.9 vs. 27.8) One cause for this may be that the 20A group size in test 2 (19 students) was significantly less the 43 subjects of test 1.

The ability of individuals to answer the same items across both tests may be an important facet of answering behaviour in Q_Lex. This statement is worth briefly examining to demonstrate why it is important. Figure 2.7 shows the relationship between student ability and item facility. Each '1' represents an item which was answered twice. The data matrix has been ordered by candidates (total score order) and items by difficulty (from top to bottom.) The matrix has an area in the upper right quadrant where most 1s accumulate. The first two students show impressive consistency in their scores (39 and 37 out of 50), and then there is large gap until the next candidate who recognized 24 items in both tests. In fact, most of the students are in the total score range of 9 to 24. Most of these recognition hits are confined to the

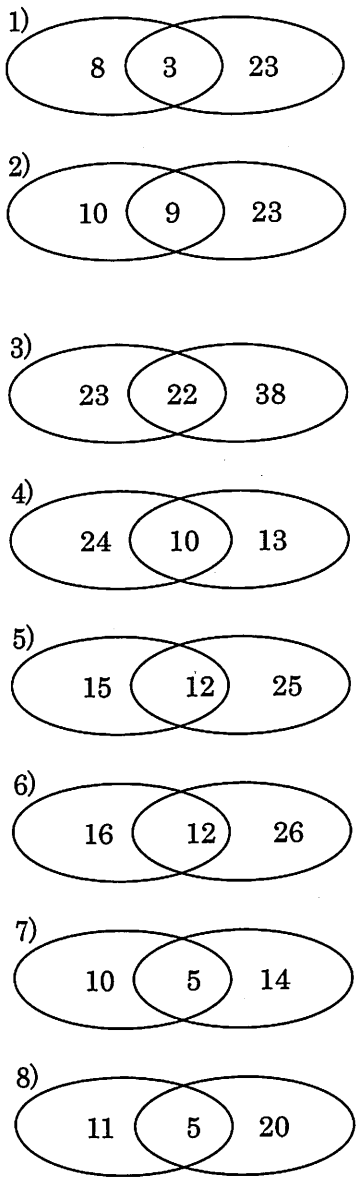
upper section of the matrix. Specifically, a line from the top right corner curving down to the bottom left shows the approximate demarcation of repeatedly recognized words. Overall, of the 533 double recognition instances in the table, about half of these (256) are accounted for by the top 12 items. In other words, there is a core of items that we would expect all students to be fairly likely to recognize consistently, if they answered them in the first test. For this reason, having a high proportion of recognition hits coming from the same word in both tests is a useful and important reflection of recognition ability, as opposed to randomly answering different items in different tests. Notwithstanding, there is a certain degree of diffuseness in the table, and the line of demarcation is not particularly well defined. Therefore, it would not be reasonable to expect extremely high consistency in repeated item recognition.

Concerning this, as figures 2.2a and 2.2b showed there is considerable variation in the consistency of students even if they have the same initial score. For example, students 24 and 26 in 1a had very similar scores (23 and 24) in the first test, and of these answered 22 and 10 again in the second test. Information on performance of individuals can be shown in the form of Venn Diagrams. These display not only the raw scores but also the overlap of items answered in both tests. The following examples from the 1OA group of students demonstrate that we must be

Figure 2.7 Ability of students plotted against facility of items.



somewhat careful in interpreting re-test consistency, and not look only at the first and second test raw scores.



In examples 1) and 2), the students both had poor initial scores, but both finished with the same, much better score in test 2. However, the degree of consistency of case 2) (9) clearly much better than that of 1) (3). Although the raw scores in both tests are very similar, a plausible case could be made that the individual in case 2) has better ability. She extended her score to 23 whilst answering almost all the items from test 1. In contrast, her counterpart was far more erratic in her answering pattern.

In examples 3) and 4), both students recorded a similar score in test 1. For an unknown reason, though, the student in example 4) performed very poorly in test 2.

Despite this, the number of items (10) which were seen in both tests was rather high, at least relative to her score in test 2.

The following examples from the 20A group of students further illustrate the usefulness of considering performance in this way.

In examples 5) and 6), there was a similar degree of performance in both tests and in the overlap of items recognized in both instances. Moreover, 12 of these items in both cases were the same. Items which were most most frequently answered in both tests are discussed below.

Conversely, in examples 7) and 8), the students did not perform very well across both tests. The students had no items in common of those they answered in the overlap.

We can also consider the whole set of students and see which items were frequently answered correctly in both administrations. 35 students took the 10A test twice. On average, they answered 10.7 items (30.6%) twice, while the 19 students who took the 20A test twice answered only 4.9 items (25.9%) twice. On this measure, the 10A items offer considerably more reliability. To understand why, we need to examine the items which performed best and worst in both kinds of masks. Tables 2.5a and 2.5b shows the items which were most and least frequently answered in both tests. 2.5a shows the 10A variety. Although the pattern is not very clear cut, it appears that the most frequently recognized items are surrounded by un-English looking

consonants clusters. This is clearly the case in numbers 1,2,3,5 and also in 4 and 6 if we consider only the left side of the target word. In the least frequently recognized items, this trend is less pronounced although number 6) [entndtsregionmw] is as obvious exception. There is no obvious reason why this was harder than [nfwmtremindvnpn].

Table 2.5a: Items answered twice in the 10A test (maximum hits: 35)

No.	Frequently recognized	# of hits	Infrequently recognized	# of hits
1)	awbrightbteoegt	31	cttememorypeewm	3
2)	nfwmtremindvnpn	27	ospotatooroeyt	2
3)	gdydinnerrbgeed	24	lnpdcetprincets	1
4)	idlswpinsistpiu	24	ldacceptlswppiu	1
5)	uocnchangemhetw	22	ocfagardenhatrw	1
6)	ntutsswitchluyn	22	entndtsregionmw	1
7)	haoanswerupmere	21	nclnaturedocfah	0

Table 2.5b shows the items from the 20A test. It is noticeable that the most reliably twice-recognized items are not surrounded by dense consonant clusters in the way that they are for the 10A items. Further, the degree of reliable recognition is lower in the 20A items where [inenchangeattet] was found in 74% of instances compared to 89% for the top 10A item. The question of what makes an easy or hard 20A item is slightly harder to answer. Some of the least recognized items, on the right-hand side, contain some minimal English words which went unnoticed before being used in the case. In number 1), for instance, 'yes' forms an integral part of the target word 'energy' . The best recognized items have almost none of these, with only number 6) having 'the' as an integral part of the target 'length'.

At any rate, on this evidence, it seems as if the most reliably recognized 10A items are commonly surrounded by dense consonant masks which are typical of first-order approximation sampling. Second-order approximation sampling is much less likely to produce consonant strings, although this does not necessarily prevent reliable recognition in multiple tests.

Table 2.5b: Items answered twice in the 20A test (maximum hits: 19)

	Frequently recognized	# of hits	Infrequently recognized	# of hits
1)	inenchangeattet	14	teffenergyesusi	1
2)	sulanswerituton	13	ugededeprinceth	1
3)	veacceptexathic	12	cechexcuseingho	1
4)	lenvepreregiont	11	eindinnerievewe	1
5)	erusimarrivemmu	11	gemeaglobaltyma	1
6)	onsulengtherryp	10	etaspiritusheyc	1
7)	disememorymaiso	10	gesilverepepryo	0

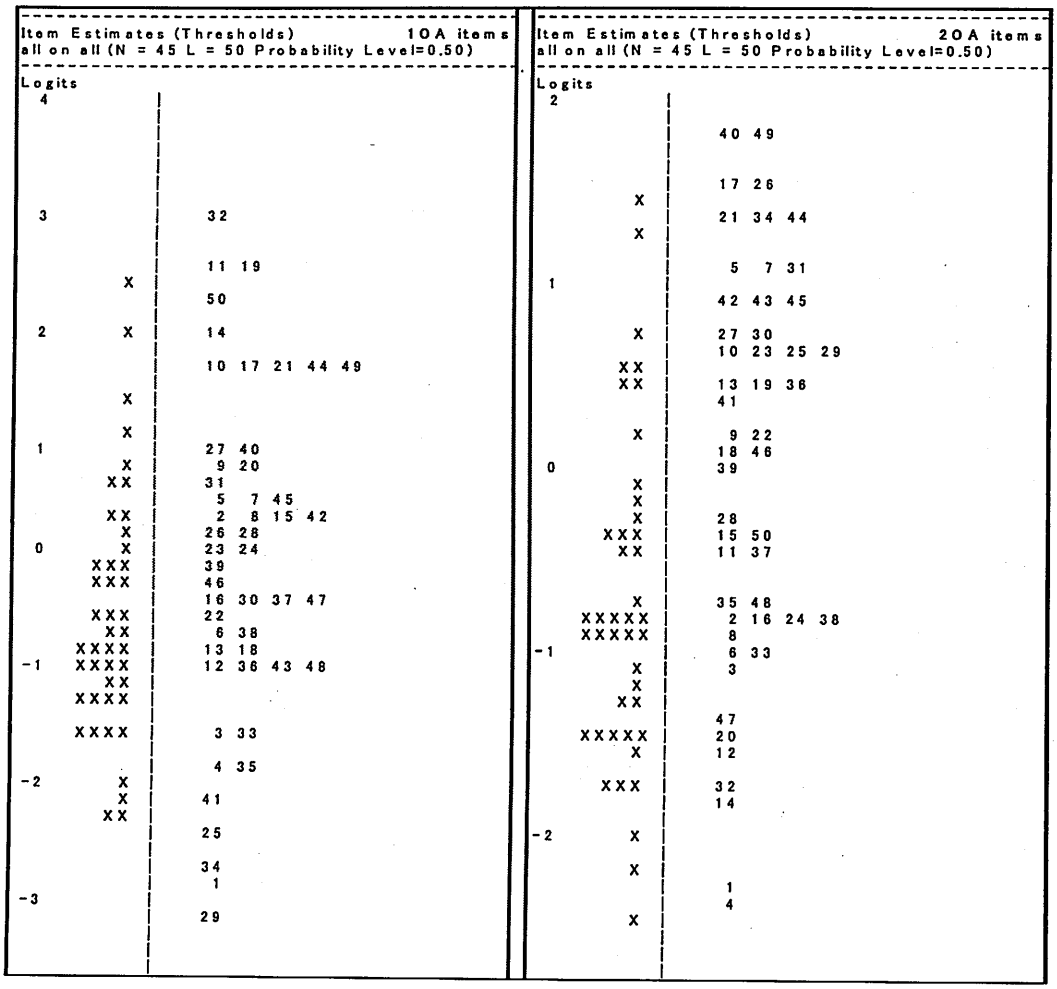
So far, I have discussed items that are reliably recognized, and what they look like. However, simple item facility (in raw scores) may not be sufficient in determining which are items are

Table 2.6: Infit Mean Square for the 10A and 20A tests					
Summary of Case Estimates			Summary of Item Estimates		
	10A	20A		10A	20A
Mean	1.00	1.00	Mean	1.00	1.00
SD	0.17	0.12	SD	0.12	0.15

best. Rasch analysis provides a method of looking at the relationship between the relative performance of all test takers on an assessment tool, such as Q_Lex, and the group of all items on the test. Further, information is provided on the difficulty of item and the ability individuals in the form of fit statistics. Table 2.6 shows that case estimates (individuals) shows slightly less variation around the mean in the 20A test, while the figures for the item estimates are almost the same. The broadly accepted range of fit for items is in the range of 0.75~1.30 (McNamara, 1996, p173). In both the 10A and 20A items, only 2 out the 50 items were outside this range. Overall, the performance of the two types of items appear comparable. With Rasch, it is possible to generate a person ability and item difficulty map (figure 2.8). This is scaled in logits (shown down the left side.) This is an interval scale, so it allows us to observe the degree of relative difficulty for each item. (The items are shown as a number from 1 to 50 on the right hand side.) The average difficulty of items is set at zero logits. Those items above 0 are above average difficulty, and vice versa. The ability of individuals is also shown on the same scale in the map. Each student is represented by an X. Rasch is a probabilistic model, and by convention testees have a 50% chance of correctly answering items which appear at the same level. For example, in the 10A map, the individual who appears at the 0 logit level had a 50% chance correctly answering items 23 and 24, and an increasingly better chance of answering the items the further they appear down the scale. Conversely, she had a smaller chance of answering the items higher up. If the item is 1 logit less than a testee's ability, she has a 75% chance of answering it, and a 90% chance if it is 2 logits below. Conversely, if the item is 1 logit higher than her ability, the chance of answering it falls to 25% and only 10% if it is 2 logits above. Finally, students with greater ability appear further up the logit scale, and vice versa. In the 10A map, the strongest individual (the X which appears at the top of the figure) had an almost 50% chance of answering the most difficulty items 11 and 19, but was almost certain of having answered those at the bottom of the scale.

On this basis, it is possible to make an informed judgment about which items are desirable to include in a revised version of the test. In Q_Lex, we are interested in the recognition ability of students on basic items of vocabulary without the mask making the item unduly hard to spot. Good performing items should therefore offer a reasonable chance of being identified, whilst leaving latitude for reflecting improved (or degraded) ability at a later point. As I reported at the beginning of Part 2, it was found that a group with one year's extra university study scored about 5 extra points on the test, compared to a matched group. To capture this change in ability change over time, it would be better not to include too many difficult items which would probably remain recognized across two administrations. Conversely, if too many easy items are included, there is less scope for observing improvement in students' scores across time. On this rationale, it may be prudent to take those items only 1 logit above and below 0 since most

Figure 2.8 Person and Item Difficulty Maps for the 10A items (left) and 20A items (right)



testees are gathered in this area of the map. In the 10A set, this would go from item numbers 27 to 48 (31 items.) In the 20A set, this would go from numbers 42 to 33 (32 items.) Putting these together would create a new set of good performing items. Excluding the items from both sets which share the same target word, this set comprises 39 items. This is slightly less than the standard 50 word set, but the shortfall could be made up from items from the original 1.0 Q_Lex version (not reported here) once the best performing items have been identified by similar procedures.

Concerning the evaluation of all individuals in the test, we have already seen in the discussion with Venn diagrams that two students with similar scores in test 1 and test 2 can show a quite different pattern of reliability in test consistency. We can also look at the degree of overall speed, measured in milliseconds, as an additional indicator of consistency on the 50 items in the test. Figures 5a and 5b showed that students with similar scores sometimes have quite dissimilar mean reaction times. This was expressed as the degree of variation in recognition speed (standard deviation of reaction time.) In 5b, five students (cases 17 to 21) scored 16 points. However, their standard deviation of reaction time varied from 7 to 26. Figure 5a also shows instances of strong

variation amongst student of similar scores. In one sense, this is not problematic since in all scoring instances the students had satisfied the basic criterion of recording a time faster than the native norm. If we look a little more closely at this, though, we can find potentially useful information in the patterning of reaction times for stronger and weaker students. Figure 4 showed the distribution of 3 students' hits in terms of their reaction times. The items which the highest scoring student recognized were not merely below the norm line of native speakers, most of them (24/38) were actually below the mean reaction time of native speakers. This is a quite marked qualitative difference from distribution of the recognition hits of the average scoring student. Here only 3 of her 17 items were below the native speaker mean reaction time line. The lowest scoring student had 2 such hits. Although somewhat arbitrary, counting the proportion of hits below the NS mean may be a further useful method of evaluating the performance of individuals. As a further example, if we look at the 5 students mentioned above from figure 5b who scored 16 points, and plot their scores against the line of native mean reaction time, we find that one student in particular has more items under the line of mean native reaction times than the other four combined and all her other items were also very near this line.

This observation will require further examination to decide if there is merit in evaluating the scores and reaction times of students in this manner.

Conclusion

This paper started by looking at three experiments that are very important for the field of word recognition. My critiques found that the tests advocated in these papers all, to varying degrees, suffer from avoidable weaknesses. For example, the test devised by Shiotsu conflates word recognition with reading ability itself, concluding that if reading is efficient, word recognition is too. This is not necessarily so. My approach was to look at vocabulary recognition time not from an interactionalist perspective, where recognition is only assessed during a reading task, but from a 'trait' perspective, which attempts to elucidate the specific roles of component processes in reading. The Q_Lex test format was designed with the aim of eliminating as many extraneous factors as possible since these often significantly impede the accurate assessment of component processes, such as basic vocabulary recognition time.

In previous research using Q_Lex, I found that students a year ahead of comparable peers in the same course of study have faster access time for the same level of vocabulary. However, reliability was a persistent concern in that experiment. To investigate this further, the experiment reported here considered the relative performance of two kinds of masking strings for the same set of 50 words used in Q_Lex. The hypothesis was that the newly constructed 20A items, which are harder to recognize, would be answered more slowly and more reliably, as reflected by the scores from two test administrations. 20A items were slower, but overall they did not prove more reliable, although for the average ability students this difference was not so strong. However, the difference was not much and many of these 20A items can be combined with 10A items and used in a proposed new set of 39 items. These particular items were judged reliable by Rasch analysis. It also seems that the masks themselves behave in rather an idiosyncratic fashion. In 10A masks, it appears that dense consonant clusters, particularly to the left of the target word aid in reliable recognition. However, in the 20A variety, it is harder to discern what causes one item to be much more reliable than another. It may be that 20A

strings can be best identified by trial and error, whereas the 10A strings can be constructed more deliberately, in future.

An important finding of this experiment was that the number of items identified in consecutive experiments is as important in measuring performance on Q_Lex as the simple raw scores from test 1 and test 2. It is this overlap which shows the degree of consistency as opposed to the more random recognition of items by students in one test or the other. I identified a set of items which have the highest chance of being recognized consistently in consecutive tests. In future research, I will investigate whether these items can reliably reflect students' recognition ability in a single test.

Further, I discussed the interesting ability of the strongest candidates to answer the some items faster than the mean reaction time for native speakers. If this ability does constitute a qualitative difference from other test takers, we may be able to evaluate test takers in a more fine-grained fashion than simply describing them with raw scores. It is also possible that mid-ranking test takers could also be discriminated in this way.

On a final note, it was, unfortunately, surprisingly difficult to spot all examples of minimal English words in the strings, especially if they are 2-letter words. For example, in eindinnerievewe the presence of 'in' went unspotted until used in the experiment. This is unfortunate, but since the students were instructed to search for 6-letter words, it is to be hoped that the presence of such minimal words has not unduly influenced test performance.

References

Bruck, M. (1990). Word-Recognition skills of adults with childhood diagnoses of dyslexia. *Developmental Psychology* 26, 3 439-454

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In Bachman, L. F. & Cohen, A.D. (eds.) (1998). *Interfaces between Second Language Acquisition and Language Testing Research.* Cambridge Applied Linguistics

Coulson, D. (2005). Recognition speed for basic vocabulary. Presentation given at the EuroSLA conference, Dubronivk, Croatia.

Goodman, K.S. (1967). Reading: A psycholinguistic guessing game. *Journal of the ReadingSpecialist*, 6, 126-135

Haynes, M.E. (1989). Individual differences in Chinese readers of English: orthography and reading. Unpublished PhD dissertation, Michigan State University.

Haynes, M. & Carr, T.H. (1990). Writing System Background and Second Language Reading: A Component Skills Analysis of English Reading by Native-Speaker-Readers of Chinese. In Carr, T.H. & Levy, B.A. (Eds.), *Reading and its Development: component skills approaches.* San Diego: Academic Press

Jacobsen, J. (1995). Word Recognition Index (WRI) as a quick screening marker of Dyslexia. *The Irish Journal of Psychology*, 16, 3, 260-266

JACET 8000英単語(2005). 桐原書店

Kamimoto, T. (2001). An examination of Nation's (1990) Vocabulary Levels Test. *Kumamoto Gakuen University*

Koda, K. (2005). *Insights into Second Language Reading: A Cross-Linguistic Approach.* Cambridge University Press

Lambert, W.E. (1990) Persistent issues in bilingualism In Harley, B,. Allen, P.,Cummins, J., Swain, M. (eds), *The Development of Second Language Proficiency Cambridge*: Cambridge University Press

Laufer, B. & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16, 1, 33-51

Laufer, B. & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? EUROSLA *Yearbook* 1(2001), 7-28.

McNamara, T.F. (1996) *Measuring Second Language Proficiency.* Pearson Education.

Meara, P. (1996). The Dimensions of Lexical Competence. In G. Brown, K. Malmkjaer and J. Williams (eds.), *Performance and Competence in Second Language Acquisition* (pp.35-53) Cambridge: Cambridge University Press.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. Language Testing 4, 142-154

Meara, P. & Wolter, B. (2004). V_Links. Beyond Vocabulary Depth Angles on the English-Speaking World 4. 85-96

Melka, F. (1997). Receptive vs. Productive vocabulary. In Schmitt and McCarthy (eds.), (1997), pp.84-102

Miller, G. A. (1963). *Language and Communication.* McGraw-Hill Company. Inc.

Nation, P. (1983). Testing and teaching vocabulary. Guidelines 5, 12-25.

Read, J. (2000). *Assessing Vocabulary.* Cambridge University Press

Richards, J. (1976). The Role of Vocabulary Teaching. *TESOL Quarterly* 10, 77-89

Rayner, K. & Balota, D.A. (1989). Parafoveal preview and lexical access during eye fixations in reading. In W.D. Marslen-Wilson (Ed.), *Lexical representation and process.* (pp.261-290). Cambridge, MA: Bradford.

Siegel, L.S. (1998). Phonological Processing Deficits and Reading Disabilities. In Metsala, J.L. & Ehri, L.C. Eds.), *Word Recognition in Beginning Literacy* (pp.141-160). Lawrence Erlbaum Associates

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of the Vocabulary Levels Test. *Language Testing* 18 (1) 55-88

Schmitt, N. and McCarthy, M. (eds.), (1997). *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge University Press

Stanovich, K.E. (1982). Individual differences in the cognitive processes of reading: 1. Word Decoding. *Journal of Learning Disabilities,* 15, 8, 485-493

Wesche, M. and Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: depth vs. breadth. *Canadian Modern Language Review* 53, 13-39

APPEXDIX 1

A New Version of Haynes & Carr's Test

DON'T TURN OVER YET

This is a word reading quiz. Please try it!

In this quiz, there are 4 kinds of questions.

On page 1, you must decide if two words have the same spelling or different spelling. First let's practice.

		SAME	DIFFERENT
went	want	S	D
long	long	S	D
same	some	S	D

On page 2, you must decide if two non-words have the same spelling or different spelling.

trup	trup	S	D
wint	wunt	S	D
dask	dosk	S	D

On page 3, you must decide if non-words with crazy spellings are the same or different.

wnyt	wnyt	S	D
lzol	lzot	S	D
verj	verj	S	D

On page 4, you must decide if two have the same *meaning*, or different meaning.

swim	walk	S	D
feel	touch	S	D
whole	half	S	D

Each page, you have 20 seconds. Please do not start the next page until I tell you.

page 1

**********DON'T LOOK YET**********

GET READY .....

			Same	Different
1	cake	cape	S	D
2	care	card	S	D
3	come	come	S	D
4	debt	debt	S	D
5	earn	earn	S	D
6	face	fact	S	D
7	fair	fear	S	D
8	free	free	S	D
9	gift	gilt	S	D
10	gold	told	S	D
11	hear	heat	S	D
12	hope	hole	S	D
13	idea	idea	S	D
14	mark	mark	S	D
15	most	mast	S	D
16	name	mane	S	D
17	park	part	S	D
18	rule	rule	S	D
19	spot	spit	S	D
20	wash	wash	S	D

page 2

**********DON'T LOOK YET**********

GET READY .....

			Same	Different
1	bele	bule	S	D
2	dirb	dirb	S	D
3	fope	fape	S	D
4	dosk	posk	S	D
5	ferm	felm	S	D
6	plad	plad	S	D
7	hige	hige	S	D
8	jume	jame	S	D
9	lipe	lipe	S	D
10	nood	dood	S	D
11	pais	pais	S	D
12	pilk	pilt	S	D
13	teal	telp	S	D
14	goot	goot	S	D
15	feem	foem	S	D
16	mign	vign	S	D
17	seaf	seaf	S	D
18	rimp	rimp	S	D
19	fain	faln	S	D
20	beal	feal	S	D

page 3

**********DON'T LOOK YET**********

GET READY .....

			Same	Different
1	botp	botp	S	D
2	bwef	bwer	S	D
3	zlok	ztok	S	D
4	dyrk	dyrk	S	D
5	drij	drij	S	D
6	fibk	gibk	S	D
7	focm	focm	S	D
8	gpal	gpal	S	D
9	hwat	nwat	S	D
10	hruv	hruv	S	D
11	jakt	javt	S	D
12	ladn	ladn	S	D
13	mukh	mukt	S	D
14	paql	poql	S	D
15	xate	xate	S	D
16	safc	mafc	S	D
17	ghop	ghop	S	D
18	ztar	ztar	S	D
19	teml	tuml	S	D
20	wnit	wnit	S	D

page 4

**********DON'T LOOK YET**********

GET READY .....

			Same	Different
1	fail	pass	S	D
2	kind	type	S	D
3	easy	hard	S	D
4	find	lose	S	D
5	like	hate	S	D
6	loss	gain	S	D
7	high	tall	S	D
8	none	many	S	D
9	keep	save	S	D
10	pull	push	S	D
11	stone	rock	S	D
12	shut	open	S	D
13	sing	talk	S	D
14	test	exam	S	D
15	soft	hard	S	D
16	poor	rich	S	D
17	sink	rise	S	D
18	pull	drag	S	D
19	warm	cool	S	D
20	swim	walk	S	D

APPENDIX 2

The Word Chain Test

Name:

1 aboveundermake	29 sameusor	57 chanceeffectcry
2 manyfighthouse	30 yetamongturn	58 thankwhilehow
3 smileownlong	31 muchwhyput	59 havefieldat
4 recordmyable	32 peoplewhitetop	60 studentwhichyoung
5 dresshotmemory	33 wifeactwar	61 housestudyway
6 itsairdeep	34 suretimeold	62 duringeasycolor
7 stationforeignfish	35 nolesshand	63 fewbuychange
8 certainbyelse	36 sitmeanlarge	64 worldgonext
9 lookshouldlow	37 pastthingback	65 underminutemind
10 remainfactside	38 besaymember	66 farexpectwait
11 kidroundbase	39 diegetcarry	67 readhistry
12 agreebadbrother	40 enoughdogour	68 servebodysun
13 modernshetell	41 plananyover	69 uselevelenter
14 amountseeleg	42 weuntilreal	70 anvalueadd
15 mouthcarnumber	43 askdoctorhalf	71 quiterestart
16 specialtryhope	44 abovewrongthough	72 morebuthead
17 allfamilylet	45 systemsofind	73 searesultwork
18 canfuturemight	46 buildfrontstory	74 shortpaydo
19 uplightnew	47 childlastnow	75 casesuchfor
20 booknewmost	48 wouldarriveme	76 soundstayname
21 formroomafter	49 walkasreceive	77 liveputdecide
22 couldanimalhappy	50 sonlistenagainst	78 himgivethese
23 ithimlie	51 leavenaturemusic	79 nightalongsocial
24 meetmoveleft	52 rulebigany	80 hardlosefire
25 makepicturedream	53 weekfightwish	81 causestarthrough
26 moneygroupneed	54 tolovecome	82 lotfromwho
27 askoldtype	55 mustbedtest	83 controleatking
28 nowbreakcity	56 stepifidea	84 bothverysign
85 highmachineforce	97 ownringpretty	109 sortperiodboy
86 themflygood	98 notesizesell	110 laughseemdrop
87 hejoinevening	99 dealrivervarious	111 greenonlycommon
88 viewsetdead	100 dayeachmay	112 birdcultureever
89 touchjobmile	101 yesrunlead	113 pickonsubject
90 freenewsnatural	102 lawtreeright	114 armtravelprovide
91 helpbestyou	103 problemredpiece	115 choosesleepaccept
92 drivebutrise	104 notbabywell	116 cutwithfloor
93 girlsimplysome	105 justdarksuggest	117 energycatchsee
94 savewindcost	106 footherfast	118 talksensewatch
95 finishraiselocal	107 evensuccesspull	119 ofwordline
96 serviceoutnice	108 publicbluemain	120 whyeyeschool

The Letter Chain Test

1 OUCCNEMHHE	28 GDYYRBGEED	55 PIOOSISTTUY
2 MQQETOEDDR	29 AWWBTEEGTJ	56 HICCTHHNSD
3 TRPPOGHEEA	30 HLPPMTANNS	57 FGGKOASGGM
4 FEFFLABNNA	31 OUUCNMHHTW	58 OTAALIISHF
5 NLLADMDDNR	32 ENNDWTSWWT	59 PFLLOSEMME
6 TNDDTXSMMW	33 ASSNBNOIIA	60 ATHAAIMFFG
7 MHHLPSSTAH	34 NFFMTRVNNP	61 SNNTFNTTAS
8 MEENRREUOI	35 IDDLSWPPIU	62 PEIITMREEN
9 HLAASLFOOPA	36 AFFENLCCBN	63 YUEEPOSSPT
10 HAOOUPMMER	37 UTTDALNEES	64 ORIIREMIIC
11 NCLDDFAHHO	38 DETKKCTTEU	65 HRAARAMMHZ
12 ASNBBNRRIA	39 PAJLLCDHHA	66 OEFOOTUUTA
13 PAJJLACCDT	40 UAFFKEQQNE	67 SHFFAEIOOF
14 DACCLSWPPI	41 NTUUTSLYYN	68 RECCRTTREV
15 ROOSOYCMMK	42 OCCATYNNAJ	69 SHIVVDTTJH
16 UREFFNACCN	43 AEAANPBBPL	70 MRRSWCEEMT
17 LUUWIAAWNA	44 LNPPDCTTSA	71 SYYSTONNLC
18 IHTTLWJEEH	45 DUCCNVOOSW	72 TNDAAUYYTE
19 OCFFAHARRW	46 CFFUETCCKT	73 ESHLLPOSSI
20 CTTEPEEWMW	47 HYYITNNCLG	74 EOOAIANTTW
21 EOTTVATAAE	48 UPMMERREFN	75 LEEMDLQQCE
22 FNEEAEAAGF	49 AWBBTEOOEG	76 AIFFTETTHR
23 EMOWWEMMOC	50 OSSPORROEY	77 COMMROVVSN
24 SIIHALWJJE	51 HLLAESPIIA	78 OTEEORISSE
25 GDYRRBGEED	52 KIMAALNNSQ	79 ANEHHFPLLF
26 OHOOSTTNBW	53 MCMMELNTTD	80 ATTPEACCEO
27 CLNNPDDETS	54 MEONNMTAAY	

APPENDIX 3:

1OA and 2OA strings and the multiple choice answers.

First order items	Second order items	Multiple choice items
uocnchangemhetw	enenchangeattet	.chunky.chance.chatty.change.
mqoertofriended	locteindesertum	.desire.desert.design.detail.
utrfamilypoghea	lolfamilypededede	.famine.famous.fasten.family.
haoanswerupmere	sulanswerítutun	.affect.answer.anyway.annual.
nladenoughmdnre	urenenoughentar	.energy.enough.engage.enigma.
entnddecidetsmw	marvedecideneme	.decade.deceit.decide.define.
hlpmanimalstahr	sereanimaloyoll	.anyhow.anthem.annual.animal.
meenreresultuoi	enasturesultyma	.resume.resign.result.resist.
hlasldoctorfota	masildoctoristi	.docile.docket.doctor.donkey.
mtafeffectlabnn	stureffectifilo	.effort.effect.engine.either.
nclnaturedocfah	aninaturenemest	.native.nation.natter.nature.
asnbnoarriveiao	erusimarrivemmu	.arrive.around.artist.arrest.
pajlchanceacdut	othichanceranoc	.chance.change.chalet.chapel.
ldacceptlswppiu	veacceptexathic	.accord.access.accept.accent.
osoroeycommontn	aysphincommonge	.common.commit.comedy.combat.
urefnatravelcnc	celttetraveltym	.tragic.travel.treble.trendy.
luwienergyawnag	teffenergyebusi	.emerge.endure.engage.energy.
ihaattacklwjeha	ithattackievesp	.attack.attend.attach.attain.
ocfagardenhatrw	emeagardengutri	.garish.garlic.garden.garret.
cttememorypeewm	disememorymaiso	.member.memory.melody.merely.
eotvefamousataa	ilypefamousutha	.famous.family.famine.fumble.
fnexpectateeaea	brexpecthreryic	.expert.expose.expect.expend.
emchoiceowemmoh	lachoicememmonc	.chorus.chosen.choppy.choice.
ihalwmarketjeha	athacomarkethas	.member.market.marvel.marble.
gdydinnerrbgeed	eindinnerievewe	.dimmer.dining.dinner.dinghy.
ohospiritstutnb	etaspiritusheyc	.spirit.splash.spider.spiral.
clnpcornerddets	spencornerncote	.cordon.corner.corset.corpse.
gdyrreducebgeed	esicassumededis	.assume.assist.assure.assess.
awbrightbteoegt	acobjecterthean	.object.obtain.oblige.oblong.
hlpminvitestanr	risninvitesutha	.invade.invent.invite.invert.
oudollarcnmhetw	nidollareryicin	.danger.dollar.during.devote.

entndtsregionmw	lenvepregiontem	.region.regret.regime.regard.
asnblengthnoiao	onsulengtherryp	.lender.length.legend.league.
nfwmtremindvnpn	gemeaglobaltyma	.gloomy.global.glamor.glance.
idlswpinsistpiu	gequreinsistico	.insist.inside.insane.insect.
fenscreenlcncbn	estscreenithadv	.scrawl.screen.scrape.scream.
utdagsignalnsea	esalysignaliosu	.singer.simple.signal.slight.
deuniquetkcteuf	equniquesithogl	.unison.unlike.unkind.unique.
paoriginjlacduh	ceweightumaveay	.weekly.weiner.weight.weapon.
uasilverfkeqnel	gesilverepepryo	.silver.silent.silken.sickly.
ntutsswitchluyn	ninativesereavi	.native.nation.nearly.napkin.
ocatexcuseynena	cechexcuseingho	.excess.excuse.exceed.except.
aeanpvalleytbpl	inthavalleyomes	.valley.valour.vacant.vanish.
lnpdcetprincets	ugededeprinceth	.priest.prissy.prison.prince.
nducnstudiovoos	vedeestudioponc	.stuffy.studio.stupid.sturdy.
fuscreametckted	fascreamibiccte	.screen.scrape.scheme.scream.
hyitncgentleltc	gemeagentleinde	.genius.gender.gentle.ginger.
upjuniormerefna	rejunioritutons	.jungle.junior.jumble.juggle.
awmusclebteoegt	nemusclentorarl	.muscle.mussel.museum.muslim.
ospotatooroeyt	cepotatoubeywan	.potato.potion.potage.potent.

APPENDIX 4

Results of investigation of students' knowledge of the test words.

I do not know the meaning of this word = N

I know this word cannot use it = X

50 students

change		nature		famous		dollar	x=2 N=1	switch	x=7
friend		arrive	x=1	expect	x=3 N=1	region	x=8 N=5	excuse	x=2 N=3
family		chance		choice	x=1	length	x=4 N=1	valley	x=8 N=20
answer		accept	x=4 N=2	market	x=1	remind	x=2 N=2	prince	
enough	x=2	common	x=5	dinner		insist	x=4 N=8	studio	x=3
decide		travel		spirit	x=3	screen		scream	x=5 N=1
animal		energy	x=4	corner	x=2	signal	x=8 N=2	gentle	
result	x=4	attack	x=4	reduce	x=3	unique	x=1	junior	x=1
doctor		garden	x=1	bright	x=6 N=1	origin	x=13 N=1	muscle	x=4 N=8
effect	x=4	memory	x=1	invite	x=2 N=4	silver	x=1 N-1	potato	